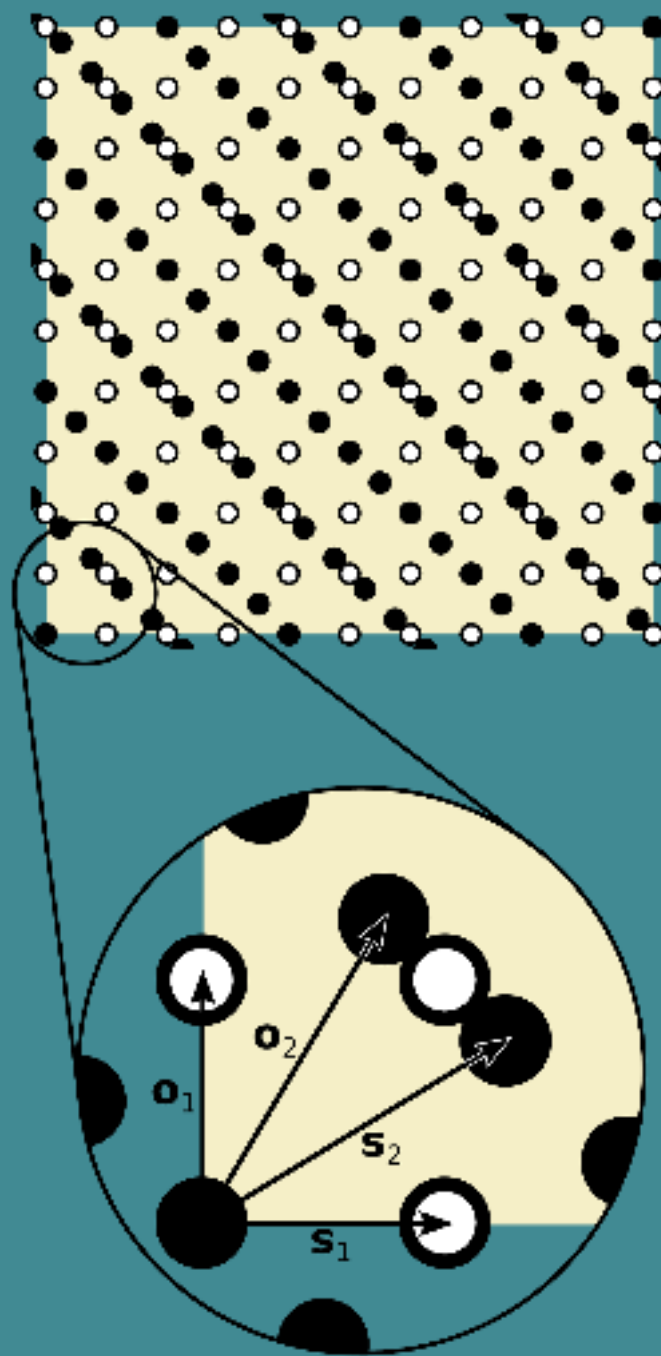


Special Relativity


Crowell



Special Relativity

Benjamin Crowell

www.lightandmatter.com

 Light and Matter
Fullerton, California
www.lightandmatter.com

Copyright ©2013 Benjamin Crowell

rev. October 18, 2017

Permission is granted to copy, distribute and/or modify this document under the terms of the Creative Commons Attribution Share-Alike License, which can be found at creativecommons.org. The license applies to the entire text of this book, plus all the illustrations that are by Benjamin Crowell. All the illustrations are by Benjamin Crowell except as noted in the photo credits or in parentheses in the caption of the figure. This book can be downloaded free of charge from www.lightandmatter.com in a variety of formats, including editable formats.

Brief Contents

1	Spacetime	11
2	Foundations (optional)	43
3	Kinematics	53
4	Dynamics	79
5	Inertia (optional)	117
6	Waves	125
7	Coordinates	143
8	Rotation (optional)	161
9	Flux	175
10	Electromagnetism	215

Contents

1 Spacetime	11
1.1 Three models of spacetime	11
Aristotelian spacetime, 12.—Galilean spacetime, 13.—Einstein's spacetime, 15.	
1.2 Minkowski coordinates	20
1.3 Measurement.	21
Invariants, 22.—The metric, 22.—The gamma factor, 25.	
1.4 The Lorentz transformation	30
1.5 Triangle and Cauchy-Schwarz inequalities.	36
Two timelike vectors, 36.—Two spacelike vectors not spanning the light cone, 37.—Two spacelike vectors spanning the light cone, 37.	
Problems	38
2 Foundations (optional)	43
2.1 Causality	43
The arrow of time, 43.—Initial-value problems, 43.—A modest definition of causality, 44.	
2.2 Flatness	45
Failure of parallelism, 45.—Parallel transport, 45.—Special relativity requires flat spacetime, 46.	
2.3 Additional postulates.	46
2.4 Other axiomatizations	48
Einstein's postulates, 48.—Maximal time, 48.—Comparison of the systems, 49.	
2.5 Lemma: spacetime area is invariant	49
Problems	51
3 Kinematics	53
3.1 How can they both ... ?	54
3.2 The stretch factor is the Doppler shift	55
3.3 Combination of velocities	57
3.4 No frame of reference moving at c	59
3.5 The velocity and acceleration vectors	60
The velocity vector, 60.—The acceleration vector, 61.—Constraints on the velocity and acceleration vectors, 62.	
3.6 Some kinematic identities.	65
3.7 The projection operator	66
3.8 Faster-than-light frames of reference?	69
3.9 Thickening of a curve	70
A geometrical interpretation of the acceleration, 70.—Bell's spaceship paradox, 71.—Deja vu, jamais vu, 73.	
Problems	75

4	Dynamics	79
4.1	Ultrarelativistic particles	79
4.2	$E=mc^2$	82
4.3	Relativistic momentum	87
	The energy-momentum vector, 87.—Collision invariants, 89.—Some examples involving momentum, 90.—Massless particles travel at c , 93.—Evidence as to which particles are massless, 94.—No global conservation of energy-momentum in general relativity, 97.	
4.4	Systems with internal structure	98
4.5	Force	100
	Four-force, 100.—The force measured by an observer, 100.—Transformation of the force measured by an observer, 102.—Work, 102.	
4.6	Two applications	103
	The Stefan-Boltzmann law, 103.—Degenerate matter, 104.	
4.7	Tachyons and FTL.	107
	A defense in depth, 107.—Experiments to search for tachyons, 109.—Tachyons and quantum mechanics, 110.	
	Problems	111
5	Inertia (optional)	117
5.1	What is inertial motion?	117
	An operational definition, 117.—Equivalence of inertial and gravitational mass, 119.	
5.2	The equivalence principle.	120
	Equivalence of acceleration to a gravitational field, 120.—Eötvös experiments, 120.—Gravity without gravity, 121.—Gravitational Doppler shifts, 121.—A varying metric, 122.	
	Problems	124
6	Waves	125
6.1	Frequency	125
	Is time's flow constant?, 125.—Clock-comparison experiments, 125.—Birdtracks notation, 126.—Duality, 127.	
6.2	Phase	127
	Phase is a scalar, 127.—Scaling, 128.	
6.3	The frequency-wavenumber covector.	128
	Visualization, 129.—The gradient, 129.	
6.4	Duality	130
	Duality in 3+1 dimensions, 130.—Change of basis, 132.	
6.5	The Doppler shift and aberration.	133
	Doppler shift, 133.—Aberration, 133.	
6.6	Phase and group velocity	136
	Phase velocity, 136.—Group velocity, 137.	
6.7	Abstract index notation	138
	Problems	142

7	Coordinates	143
7.1	An example: accelerated coordinates.	143
7.2	Transformation of vectors.	145
7.3	Transformation of the metric	146
7.4	Summary of transformation laws.	148
7.5	Inertia and rates of change	150
7.6	Volume, orientation, and the Levi-Civita tensor	151
	Volume, 151.—Orientation, 153.—The 3-volume covector, 156.	
	Problems	160
8	Rotation (optional)	161
8.1	Rotating frames of reference	161
	No clock synchronization, 161.—Rotation is locally detectable, 162.— The Sagnac effect, 162.—A rotating coordinate system, 163.	
8.2	Angular momentum	165
	The relativistic Bohr model, 165.—The angular momentum tensor, 167.	
8.3	Boosts and rotations.	170
	Rotations, 170.—Boosts, 171.—Thomas precession, 171.	
	Problems	174
9	Flux	175
9.1	The current vector.	175
	Current as the flux of charged particles, 175.—Conservation of charge, 178.	
9.2	The stress-energy tensor	179
	Conservation and flux of energy-momentum, 179.—Symmetry of the stress-energy tensor, 179.—Dust, 180.—Rank-2 tensors and their transformation law, 180.—Pressure, 182.—A perfect fluid, 182.—Two simple examples, 184.—Energy conditions, 186.	
9.3	Gauss's theorem	188
	Integral conservation laws, 188.—A simple form of Gauss's theo- rem, 188.—The general form of Gauss's theorem, 189.—The energy- momentum vector, 191.—Angular momentum, 193.	
9.4	The covariant derivative	193
	Comma, semicolon, and birdtracks notation, 196.—Finding the Christoffel symbol from the metric, 196.—The geodesic equation, 197.	
9.5	Congruences, expansion, and rigidity	201
	Congruences, 201.—Expansion and rigidity, 202.—Caustics, 204.— The Herglotz-Noether theorem in 1+1 dimensions, 205.—Bell's spaceship paradox revisited, 206.	
9.6	Units of measurement for tensors	207
9.7	Notations for tensors.	210
	Concrete index notation, 210.—Coordinate-independent notation, 210.—Cartan notation, 211.—Index-free notation, 212.—Incompatibility of Cartan and index-free notation with dimensional analysis, 212.	
	Problems	214

10 Electromagnetism **215**

10.1 Relativity requires magnetism 215

10.2 Fields in relativity. 216

Time delays in forces exerted at a distance, 216.—Fields carry energy., 216.—Fields must have transformation laws, 217.

10.3 Electromagnetic fields 218

The electric field, 218.—The magnetic field, 218.—The electromagnetic field tensor, 219.—What about gravity?, 221.

10.4 Transformation of the fields 221

10.5 Invariants 224

10.6 Stress-energy tensor of the electromagnetic field . . . 226

10.7 Maxwell's equations 230

Statement and interpretation, 230.—Experimental support, 231.—Incompatibility with Galilean spacetime, 231.—Not manifestly relativistic in their original form, 231.—Lorentz invariance, 233.

Problems 237

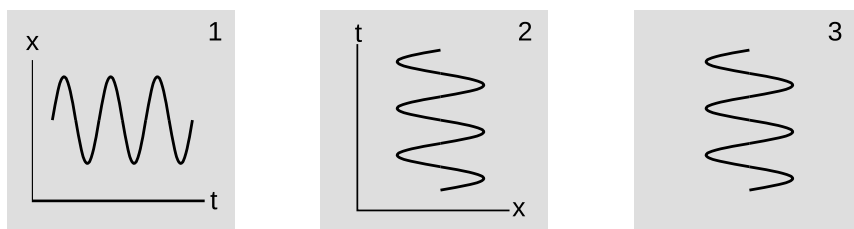
Appendix ??: Hints and solutions ??

Chapter 1

Spacetime

1.1 Three models of spacetime

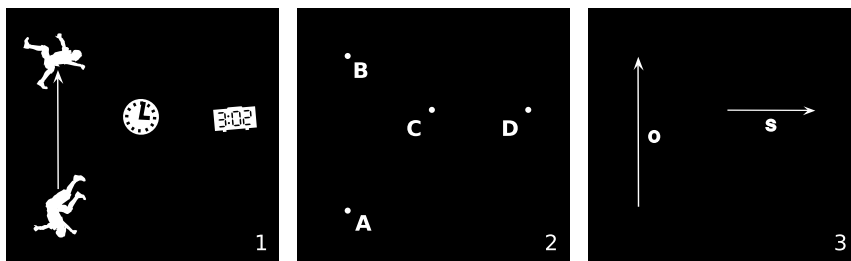
“The test of a first-rate intelligence is the ability to hold two opposing ideas in mind at the same time and still retain the ability to function.” —F. Scott Fitzgerald



a / Three views of spacetime. 1. A typical graph of a particle's motion: an oscillation. 2. In relativity, it's customary to swap the axes, and 3 we can even remove the axes entirely.

Time and space together make spacetime, figure a, the stage on which physics is played out. Until 1905, physicists were trained to accept two mutually contradictory theories of spacetime. I'll call these the Aristotelian and Galilean views, although my colleagues from that era would have been offended to be accused of even partial Aristotelianism.

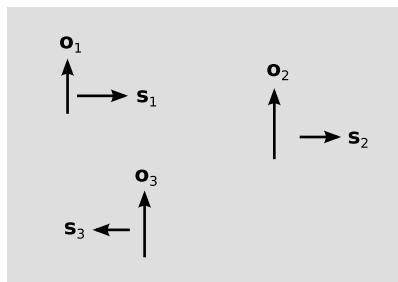
b / 1. An observer and two clocks.
 2. Idealization as events.
 3. Vectors used to represent relationships between events.



1.1.1 Aristotelian spacetime

Figure b/1 shows an observer and two clocks, represented using the graphical conventions of figure a/3. The existence of such a material object at a certain place and time constitutes an *event*, which we idealize as a point, b/2. Spacetime consists of the set of all events. As time passes, a physical object traces out a continuous curve, a set of events known in relativistic parlance as its *world-line*. Since paper and computer screens are two-dimensional, the drawings only represent one dimension of space plus one dimension of time, which in relativity we call “1+1 dimensions.” The real universe has three spatial dimensions, so real spacetime has 3+1 dimensions. Most, but not all, of the interesting phenomena in special relativity can be understood in 1+1 dimensions, so whenever possible in this book I will draw 1+1-dimensional figures without apology or explanation.

The relativist’s attitude is that events and relationships between events are primary, while coordinates such as x and t are secondary and possibly irrelevant. Coordinates let us attach labels like (x, t) to points, but this is like God asking Adam to name all the birds and animals: the animals didn’t care about the names. Figure b/3 shows the use of vectors to indicate relationships between points. Vector \mathbf{o} is an observer-vector, connecting two points on the world-line of the person. It points from the past into the future. The vector \mathbf{s} connecting the two clocks is a vector of simultaneity. The clocks have previously been synchronized side by side, and if we assume that transporting them to separate locations doesn’t disrupt them, then the fact that both clocks read two minutes after three o’clock tells us that the two events occur at the same time.



c / Valid vectors representing observers and simultaneity, according to the Aristotelian model of spacetime.

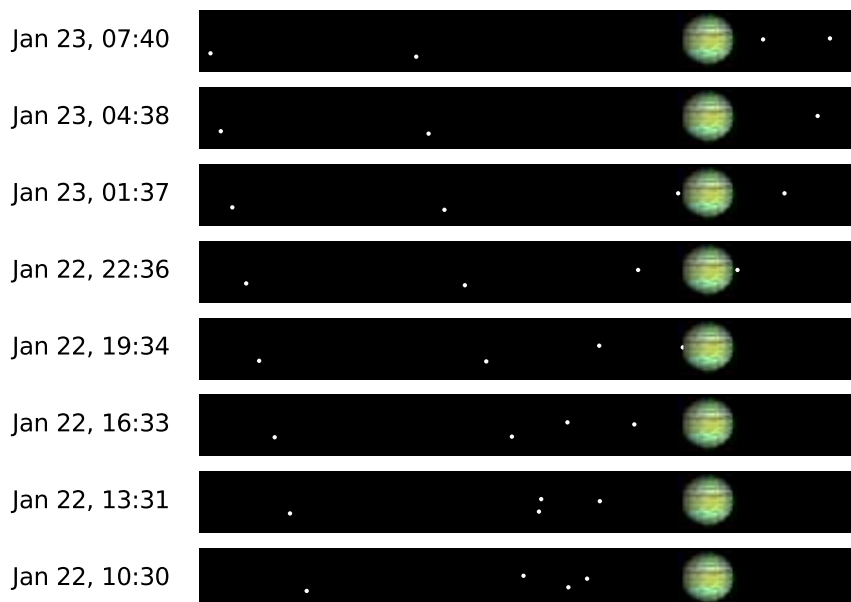
The Aristotelian model of spacetime is characterized by a set of rules about what vectors are valid observer- and simultaneity-vectors. We require that every \mathbf{o} vector be parallel to every other, and likewise for \mathbf{s} vectors. But, as is usual with vectors, we allow the arrow to be drawn anywhere without considering the different locations to have any significance; that is, our model of spacetime doesn’t allow different regions to have different properties.

When Einstein was a university student, these rules (phrased differently) were the ones he was taught to use in describing electricity and magnetism. He later recalled imagining himself on a motorcy-

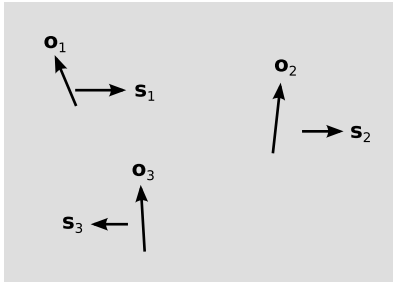
cle, riding along next to a light wave and trying to imagine how his observations could be reconciled with Maxwell's equations. I don't know whether he was ever brave enough to describe this daydream to his professors, but if he had, their answer would have been essentially that his hypothetical \mathbf{o} vector was illegal. The good \mathbf{o} vectors were thought to be the ones that represented an observer at rest relative to the ether, a hypothetical all-pervasive medium whose vibrations were electromagnetic waves. However silly this might seem to us a hundred years later, it was in fact strongly supported by the evidence. A vast number of experiments had verified the validity of Maxwell's equations, and it was known that if Maxwell's equations were valid in coordinates (x, t) defined by an observer \mathbf{o} , they would become invalid under the transformation $(x', t') = (x + vt, t)$ to coordinates defined by an observer \mathbf{o}' in motion at velocity v relative to \mathbf{o} .

1.1.2 Galilean spacetime

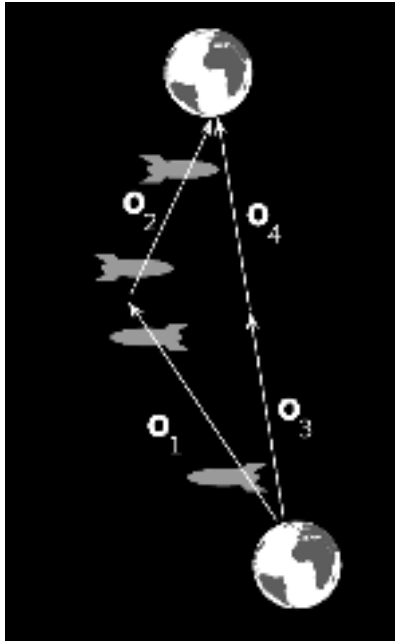
But the Aristotelian model was already known to be wrong when applied to material objects. The classic empirical demonstration of this fact came around 1610 with Galileo's discovery of four moons orbiting Jupiter, figure d. Aristotelianism in its ancient form was originally devised as an explanation of why objects always seemed to settle down to a natural state of rest according to an observer standing on the earth's surface. But as Jupiter flew across the heavens, its moons circled around it, without showing any natural tendency to fall behind it like a paper cup thrown out the window of a car. Just as an observer \mathbf{o}_1 standing on the earth would consider the earth to be at rest, \mathbf{o}_2 hovering in a balloon at Jupiter's cloudtops would say that the jovian clouds represented an equally "natural" state of rest.



d / A simulation of how Jupiter and its moons might appear at intervals of three hours through a telescope. Because we see the moons' circular orbits edgewise, their world-lines appear sinusoidal. Over this time period, the innermost moon, Io, completes half a cycle.



e / Valid vectors representing observers and simultaneity, according to the Galilean model of spacetime.



f / Example 1.

We are thus led to a different, Galilean, set of rules for \mathbf{o} and \mathbf{s} vectors. All \mathbf{s} vectors are parallel to one another, but *any* vector that is not parallel to an \mathbf{s} vector is a valid \mathbf{o} vector. (We may wish to require that it point into the future rather than the past, but Newton’s laws are symmetric under time-reversal, so this is not strictly necessary.)

Galilean spacetime, unlike Aristotelian spacetime, has no universal notion of “same place.” I can drive to Gettysburg, Pennsylvania, and stand in front of the brass plaque that marks the site of the momentous Civil War battle. But am I really in the same place? An observer on another planet would say that our planet had moved through space since 1863.

Note that our geometrical description includes a notion of parallelism, but not of angular measure. We don’t know or care whether the “angle” between an \mathbf{s} and an \mathbf{o} is 90 degrees. One represents a distance, while the other represents an interval of time, and we can’t define the angle between a distance and a time. The same was true in the Aristotelian model; the vectors in figure c were drawn perpendicular to one another simply as a matter of convention, but any other angle could have been used.

The Galilean twin paradox

Example 1

Alice and Betty are identical twins. Betty goes on a space voyage, traveling away from the earth along vector \mathbf{o}_1 and then turning around and coming back on \mathbf{o}_2 . Meanwhile, Alice stays on earth. Because this is an experiment involving material objects, and the conditions are similar to those under which Galilean relativity has been repeatedly verified by experiment, we expect the results to be consistent with Galilean relativity’s claim that motion is relative. Therefore it seems that it should be equally valid to consider Betty and the spaceship as having been at rest the whole time, while Alice and the planet earth traveled away from the spaceship along \mathbf{o}_3 and then returned via \mathbf{o}_4 . But this is not consistent with the experimental results, which show that Betty undergoes a violent acceleration at her turnaround point, while Alice and the other inhabitants of the earth feel no such effect.

The paradox is resolved by realizing that Galilean relativity defines unambiguously whether or not two vectors are parallel. It’s true that we could fix a frame of reference in which \mathbf{o}_1 represented the spaceship staying at rest, but \mathbf{o}_2 is not parallel to \mathbf{o}_1 , so in this frame we still have a good explanation for why Betty feels an acceleration: she has gone from being at rest to being in motion.

Regardless of which frame of reference we pick, and regardless of whether we even fix a frame of reference, \mathbf{o}_3 and \mathbf{o}_4 are parallel to one another, and this explains why Alice feels no effect.

1.1.3 Einstein's spacetime

We have two models of spacetime, neither of which is capable of describing all the phenomena we observe. Because of the relatively crude state of technology *ca.* 1900, it required considerable insight for Einstein to piece together a fragmentary body of indirect evidence and arrive at a consistent and correct model of spacetime. Today, the evidence is part of everyday life. For example, every time you use a GPS receiver, you're using Einstein's theory of relativity. Somewhere between 1905 and today, technology became good enough to allow conceptually *simple* experiments that students in the early 20th century could only discuss in terms like "Imagine that we could. . ."

A good jumping-on point is 1971. In that year, J.C. Hafele and R.E. Keating brought atomic clocks aboard commercial airliners, figure g, and went around the world, once from east to west and once from west to east. Hafele and Keating observed that there was a discrepancy between the times measured by the traveling clocks and the times measured by similar clocks that stayed home at the U.S. Naval Observatory in Washington.¹ The east-going clock lost time, ending up off by -59 ± 10 nanoseconds, while the west-going one gained 273 ± 7 ns.

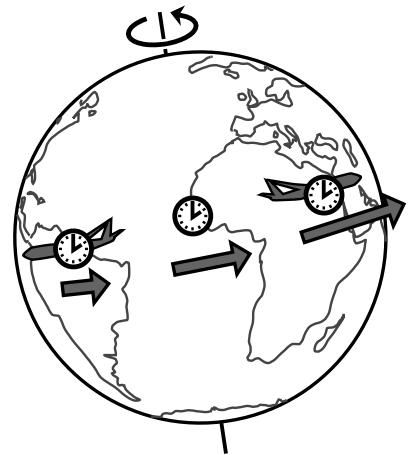
We are used to thinking of time as absolute and universal, so it is disturbing to find that it can flow at a different rates for different observers. Nevertheless, the effects that Hafele and Keating observed were small. This makes sense: Galilean relativity had already been thoroughly verified for material objects such as clocks, planets, and airplanes, so a new theory like Einstein's had to agree with Galileo's to a good approximation, within the Galilean theory's realm of applicability. This requirement of backward-compatibility is known as the correspondence principle.

It's also reassuring that the effects on time were small compared to the three-day lengths of the plane trips. There was therefore no opportunity for paradoxical scenarios such as one in which the east-going experimenter arrived back in Washington before he left and then convinced himself not to take the trip. A theory that maintains this kind of orderly relationship between cause and effect is said to satisfy causality.²

Hafele and Keating were testing specific quantitative predictions of relativity, and they verified them to within their experiment's error bars. Let's work backward instead, and inspect the empirical results for clues as to how time works. The disagreements among the clocks suggest that simultaneity is not absolute: different ob-



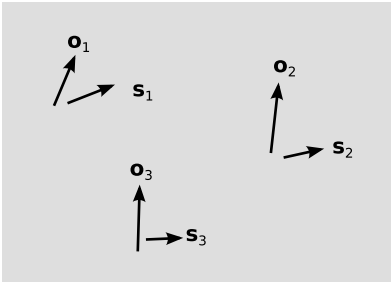
g / The clock took up two seats, and two tickets were bought for it under the name of "Mr. Clock."



h / All three clocks are moving to the east. Even though the west-going plane is moving to the west relative to the air, the air is moving to the east due to the earth's rotation.

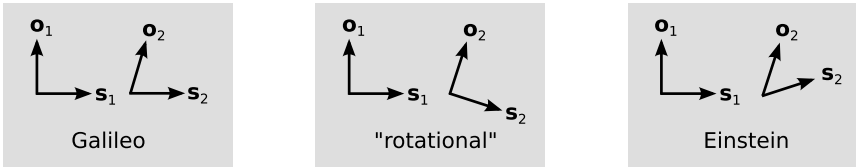
¹There were actually several effects at work, but these details do not affect the present argument, which only depends on the fact that there is no absolute time. See p. 122 for more on this topic.

²For more about causality, see section 2.1, p. 43.



i / According to Einstein, simultaneity is relative, not absolute.

servers have different notions of simultaneity, as suggested in figure i. Just as Galilean relativity freed the \mathbf{o} vectors from the constraint of being parallel to one another, Einstein frees the \mathbf{s} vectors. Galileo made “same place” into an ambiguous concept, while Einstein did the same with “simultaneous.” But because a particular observer does have methods of synchronizing clocks (e.g., Einstein synchronization, example 4, p. 18), the definition of simultaneity isn’t completely arbitrary. For each \mathbf{o} vector we have a corresponding \mathbf{s} vector, which represents that observer’s opinion as to what constitutes simultaneity. Because the convention on a Cartesian $x-t$ graph is to draw the axes at right angles to one another, we refer to such a pair of vectors as orthogonal, but the word is not to be interpreted literally, since we can’t define an actual angle between a time interval and a spatial displacement.



j / Possibilities for the behavior of orthogonality.

What, then, are the rules for orthogonality? Figure j shows three possibilities. In each case, we have an initial pair of vectors \mathbf{o}_1 and \mathbf{s}_1 that we assume are orthogonal, and we then draw a new pair \mathbf{o}_2 and \mathbf{s}_2 for a second observer who is in motion relative to the first. The Galilean case, where \mathbf{s}_2 remains parallel to \mathbf{s}_1 , has already been ruled out. The second case is the one in which \mathbf{s} rotates in the same direction as \mathbf{o} . This one is forbidden by causality, because if we kept on rotating, we could eventually end up rotating \mathbf{o} by 180 degrees, so by a continued process of acceleration, we could send an observer into a state in which her sense of time was reversed. We are left with only one possibility for Einstein’s spacetime, which is the one in which a clockwise rotation of \mathbf{o} causes a counterclockwise rotation of \mathbf{s} , like closing a pair of scissors.

Now there is a limit to how far this process can go, or else the \mathbf{s} and \mathbf{o} would eventually lie on the same line. But this is impossible, for a valid \mathbf{s} vector can never be a valid \mathbf{o} , nor an \mathbf{o} a valid \mathbf{s} . Such a possibility would mean that an observer would describe two different points on his own world-line as simultaneous, but an observer for whom no time passes is not an observer at all, since observation implies collecting data and then being able to remember it at some later time. We conclude that there is a diagonal line that forms the boundary between the set of possible \mathbf{s} vectors and the set of valid \mathbf{o} vectors. This line has some slope, and the inverse of this slope corresponds to some velocity, which is apparently a universal

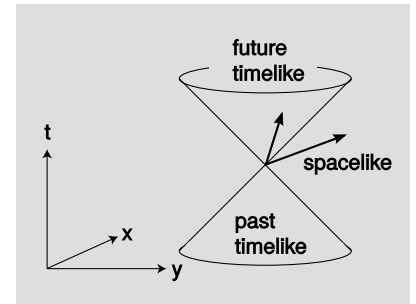
and fixed property of Einstein’s spacetime. This velocity we call c , and the correspondence principle tells us that c must be very large, because otherwise Einsteinian, or “relativistic,” effects such as time distortion would have been large even for motion at everyday speeds; in the Hafele-Keating experiment they were quite small, even at the high speed of a passenger jet.

Although c is a large number when expressed in meters per second, for convenience in relativity we will always choose units such that $c = 1$. The boundary between **s** and **o** vectors then appears on spacetime diagrams as a diagonal line at ± 45 degrees. In more than one spatial dimension, this boundary forms a cone, figure k, and for reasons that will become more clear in a moment, this cone is called the light cone. Vectors lying inside the light cone are referred to as timelike, those outside as spacelike, and those on the cone itself as lightlike or null.

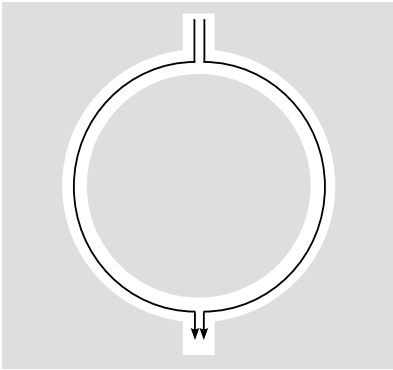
An important advantage of Einstein’s relativity over Galileo’s is that it is compatible with the empirical observation that some phenomena travel at a certain fixed speed. Light travels at a fixed speed, and so do other phenomena such as gravitational waves (first directly detected in 2016). So do all massless particles (subsection 4.3.4). This fixed speed is c , and all observers agree on it. In 1905, the only phenomenon known to travel at c was light, so c is usually described as the “speed of light,” but from the modern point of view it functions more as a kind of conversion factor between our units of measurement for time and space. It is a property of spacetime, not a property of light.

More fundamentally, c is the maximum speed of cause and effect. If we could propagate cause and effect, e.g., by transmitting a signal, at a speed greater than c , then the following argument shows that we would be violating either causality or the principle that motion is relative. If a signal could be propagated at a speed greater than c , then the vector **r** connecting the cause and the effect would be spacelike. By opening and closing the “scissors” of figure j, we can always find an observer **o** who considers **r** to be a vector of simultaneity. Thus if faster-than-light propagation is possible, then instantaneous propagation is possible, at least for some observer. Since motion is relative, this must be possible for all observers, regardless of their state of motion. Therefore *any* spacelike vector is one along which we can send a signal. But by adding two spacelike vectors we can make a vector lying in the past timelike light cone, so by relaying the signal we could send a message into the past, violating causality.

In interpreting this argument, note that neither the relativity of motion nor causality is a logical necessity; they are both just generalizations based on a body of evidence. For more on causality, and its uncertain empirical status, see section 2.1, p. 43.



k / The light cone.



I / A ring laser gyroscope.

The ring laser gyroscope

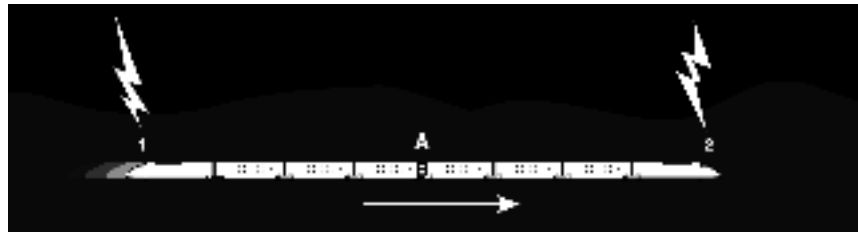
Example 2

If you've flown in a jet plane, you can thank relativity for helping you to avoid crashing into a mountain or an ocean. Figure I shows a standard piece of navigational equipment called a ring laser gyroscope. A beam of light is split into two parts, sent around the perimeter of the device, and reunited. Since the speed of light is constant, we expect the two parts to come back together at the same time. If they don't, it's evidence that the device has been rotating. The plane's computer senses this and notes how much rotation has accumulated.

No frequency-dependence

Example 3

Relativity has only one universal speed, so it requires that all light waves travel at the same speed, regardless of their frequency and wavelength. Presently the best experimental tests of the invariance of the speed of light with respect to wavelength come from astronomical observations of gamma-ray bursts, which are sudden outpourings of high-frequency light, believed to originate from a supernova explosion in another galaxy. One such observation, in 2009,³ found that the times of arrival of all the different frequencies in the burst differed by no more than 2 seconds out of a total time in flight on the order of ten billion years!



Einstein's train

Example 4

▷ The figure shows a famous thought experiment devised by Einstein. A train is moving at constant velocity to the right when bolts of lightning strike the ground near its front and back. Alice, standing on the dirt at the midpoint of the flashes, observes that the light from the two flashes arrives simultaneously, so she says the two strikes must have occurred simultaneously. Bob, meanwhile, is sitting aboard the train, at its middle. He passes by Alice at the moment when Alice later figures out that the flashes happened. Later, he receives flash 2, and then flash 1. He infers that since both flashes traveled half the length of the train, flash 2 must have occurred first. How can this be reconciled with Alice's belief that the flashes were simultaneous?

▷ Figure n shows the corresponding spacetime diagram. It seems paradoxical that Alice and Bob disagree on simultaneity, but this is

³<http://arxiv.org/abs/0908.1832>

only because we have an ingrained prejudice in favor of Galilean relativity. Alice's method of determining that 1 and 2 were simultaneous is valid, and is known as Einstein synchronization. The dashed line connecting 1 and 2 is orthogonal to Alice's world-line. But Bob has a different opinion about what constitutes simultaneity. The slanted dashed line is orthogonal to his world-line. According to Bob, 2 happened before the time represented by this line, 1 after.

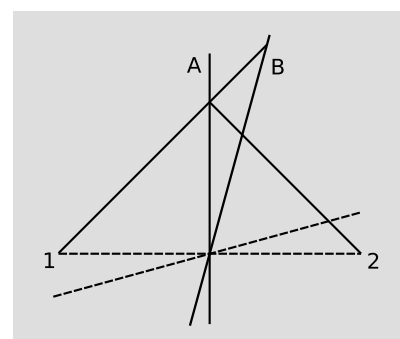
Example 4 is of course impractical as described, since real trains don't travel at speeds anywhere near c relative to the dirt. We say that their speeds are "nonrelativistic." Because Einstein coined the term "relativity," and his version of relativity superseded Galileo's, the unmodified word is normally understood to refer to Einsteinian relativity. A physicist who studies Einstein-relativity is a relativist. A material object moving at a speed very close to c is described as ultrarelativistic. One often hears laypeople describing relativity in terms of certain effects that would happen "if you went at the speed of light." In fact, as we'll see in ch. 3 and 4, it is not possible to accelerate material objects to c , and in any case that isn't necessary. Relativistic effects exist at all speeds, but they're weak at speeds small compared to c .

Numerical value of c

In this book we'll use units in which $c = 1$. However, many beginners are vexed by the question of why c has the particular value it does in a given system of units such as the SI. Related to this is the question of whether c could ever change, so that measuring it today and measuring it tomorrow would give slightly different results. In a system of units where c has units, its value is what it is only because of our choice of units, and there is no meaningful way to test whether it changes.

Let's take the SI as an example of a system of units. The SI was originally set up so that the meter and the second were defined in terms of properties of our planet. The meter was one forty-millionth of the earth's circumference, and the second was $1/86,400$ of a mean solar day. Thus when we express c as 3×10^8 m/s, we are basically specifying the factor by which c exceeds the speed at which a point on the equator goes around the center of the earth (with additional conversion factors of 40,000,000 and 86,400 thrown in). Since the properties of our planet are accidents of the formation of the solar system, there is no physical theory that can tell us why c has this value in the original French-Revolutionary version of the SI.

The base units of the SI were redefined over the centuries. Today, the second is defined in terms of an atomic standard, and the meter is defined as $1/299,792,458$ of a light-second. Therefore c has a defined value of exactly 299,792,458 m/s. Again, we find that the numerical value of c has no fundamental significance; it is merely a



n / Example 4.

matter of definition.

It is possible to form the unitless ratio $\alpha = e^2/\hbar c \approx 1/137$, called the fine structure constant. Its value does not depend on our choice of units, so it is possible to do experiments to look for changes in its value over time, e.g., by comparing the spectrum of hydrogen on earth with the spectrum of distant stars, whose light has taken billions of years to get to us. Claims have even been made to the effect that these observations do show a change, although this appears to have been a mistake. If such a change did occur, we would not be able to attribute it unambiguously to a change in c rather than a change in \hbar or the fundamental charge.

The standards used to define our units could change over time. The platinum-iridium standard for the kilogram in Paris is suspected to have lost about 50 μg of mass over the last century. Even the atomic standard used to define the second could be changing due to physics beyond our present knowledge. A change in c might produce such a change, but any such change could also be produced by changes in other physical constants, such as the others occurring in the fine structure constant. Such issues are discussed at greater length in section 9.6, p. 207.

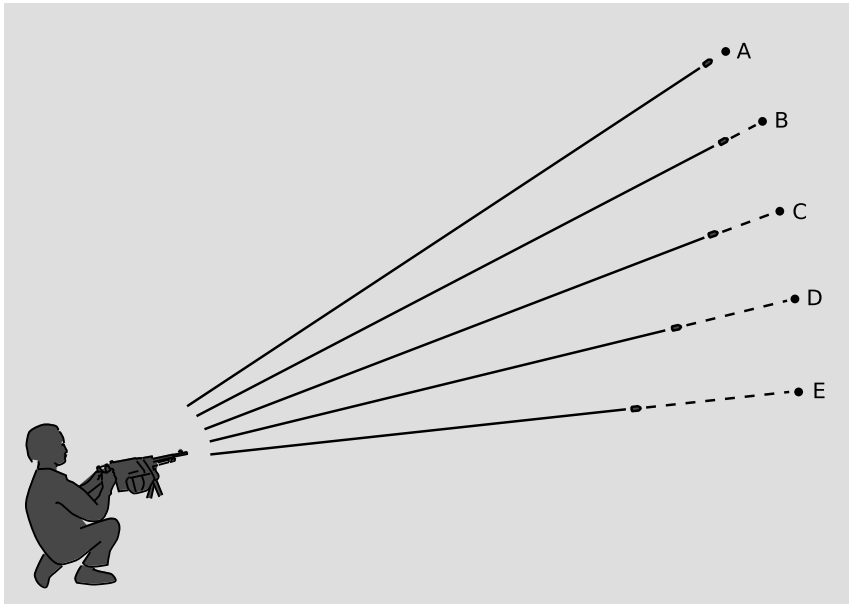
Discussion question

A The machine-gunner in the figure sends out a spray of bullets. Suppose that the bullets are being shot into outer space, and that the distances traveled are trillions of miles (so that the human figure in the diagram is not to scale). After a long time, the bullets reach the points shown with dots which are all equally far from the gun. Their arrivals at those points are events A through E, which happen at different times. The chain of impacts extends across space at a speed greater than c . Does this violate special relativity?

1.2 Minkowski coordinates

It is often convenient to name points in spacetime using coordinates, and a particular type of naming, chosen by Einstein and Minkowski, is the default in special relativity. I'll refer to the coordinates of this system as Minkowski coordinates, and they're what I have in mind throughout this book when I use letters like t and x (or variations like x' , t_o , etc.) without further explanation. To define Minkowski coordinates in 1 + 1 dimensions, we need to pick (1) an event that we consider to be the origin, $(t, x) = (0, 0)$, (2) an observer-vector \mathbf{o} , and (3) a side of the observer's world-line that we will call the positive x side, and draw on the right in diagrams. The observer is required to be inertial,⁴ so that by repeatedly making copies of \mathbf{o}

⁴For now we appeal to the freshman mechanics notion of "inertial." A better relativistic definition, which differs from the Newtonian one, is given in ch. 5.



Discussion question A.

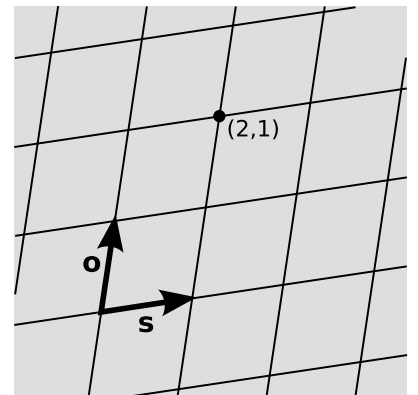


p / Hermann Minkowski (1864-1909).

and laying them tip-to-tail, we get a chain that lies on top of the observer's world-line and represents ticks on the observer's clock.

Minkowski coordinates use units with $c = 1$. Explicitly, we define the unique vector \mathbf{s} that is orthogonal to \mathbf{o} , points in the positive direction, and has a length of one clock-tick. In practical terms, the orthogonality could be defined by Einstein synchronization (example 4, p. 18), and the length by arranging that a radar echo travels to the tip of \mathbf{s} and back in two ticks.

We now construct a graph-paper lattice, figure q, by repeating the vectors \mathbf{o} and \mathbf{s} . This grid defines a name (t, x) for each point in spacetime.



q / A set of Minkowski coordinates.

1.3 Measurement

We would like to have a general system of measurement for relativity, but so far we have only an incomplete patchwork. The length of a timelike vector can be defined as the time measured on a clock that moves along the vector. A spacelike vector has a length that is measured on a ruler whose motion is such that in the ruler's frame of reference, the vector's endpoints are simultaneous. But there is no third measuring instrument designed for the purpose of measuring lightlike vectors.

Nor do we automatically get a complete system of measurement just by having defined Minkowski coordinates. For example, we don't yet know how to find the length of a timelike vector such as $(\Delta t, \Delta x) = (2, 1)$, and we suspect that it will be not equal 2, since the Hafele-Keating experiment tells us that a clock undergoing the motion represented by $\Delta x = 1$ will probably not agree with a

clock carried by the observer whose clock we used in defining these coordinates.

1.3.1 Invariants

The whole topic of measurement is apt to be confusing, because the shifting landscape of relativity makes us feel as if we've walked into a Salvador Dali landscape of melting pocket watches. A good way to regain our bearings is to look for quantities that are *invariant*: they are the same in all frames of reference. A Euclidean invariant, such as a length or an angle, is one that doesn't change under rotations: all observers agree on its value, regardless of the orientations of their frames of reference. For a relativistic invariant, we require in addition that observers agree no matter what state of motion they have. (A transformation that changes from one inertial frame of reference to another, without any rotation, is called a boost.)

Electric charge is a good example of an invariant. Electrons in atoms typically have velocities of 0.01 to 0.1 (in our relativistic units, where $c = 1$), so if an electron's charge depended on its motion relative to an observer, atoms would not be electrically neutral. Experiments have been done⁵ to test this to the phenomenal precision of one part in 10^{21} , with null results.

A vector can never be an invariant, since it changes direction under a rotation. (Some vectors, such as velocities, also change under a boost.) In freshman mechanics, any quantity, such as energy, that wasn't a vector usually fell into the category we referred to as scalars. In relativity, however, the term "scalar" has a much more restrictive definition, which we'll discuss in section 6.2.1, p. 127.

By the way, beginners in relativity sometimes get confused about invariance as opposed to conservation. They are not the same thing, and neither implies the other. For example, momentum has a direction in space, so it clearly isn't invariant — but we'll see in section 4.3 that there is a relativistic version of the momentum vector that is conserved. As in Newtonian mechanics, we don't care if all observers agree on the momentum of a system — we only care that the *law* of momentum conservation is valid and has the same form in all frames. Conversely, there are quantities that are invariant but not conserved, mass being an example.

1.3.2 The metric

Area in $1 + 1$ dimensions is also an invariant, as proved on p. 49. The invariance of area has little importance on its own, but it provides a good stepping stone toward a relativistic system of measurement. Suppose that we have events A (Charles VII is restored to

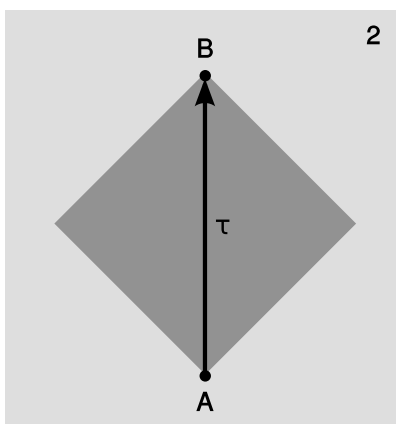
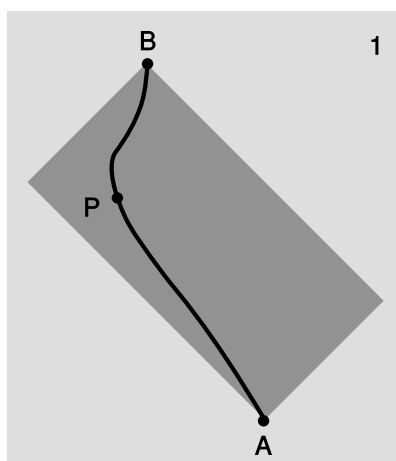


Figure 1.3.2: The two light-rectangles have the same area.

⁵Marinelli and Morigio, "The electric neutrality of matter: A summary," Physics Letters B137 (1984) 439.

the throne) and B (Joan of Arc is executed). Now imagine that technologically advanced aliens want to be present at both A and B, but in the interim they wish to fly away in their spaceship, be present at some other event P (perhaps a news conference at which they give an update on the events taking place on earth), but get back in time for B. Since nothing can go faster than c (which we take to equal 1), P cannot be too far away. The set of all possible events P forms a rectangle, figure r/1, in the $1+1$ -dimensional plane that has A and B at opposite corners and whose edges have slopes equal to ± 1 . We call this type of rectangle a light-rectangle.

The area of this rectangle will be the same regardless of one's frame of reference. In particular, we could choose a special frame of reference, panel 2 of the figure, such that A and B occur in the same place. (They do not occur at the same place, for example, in the sun's frame, because the earth is spinning and going around the sun.) Since the speed $c = 1$ is the same in all frames of reference, and the sides of the rectangle had slopes ± 1 in frame 1, they must still have slopes ± 1 in frame 2. The rectangle becomes a square, whose diagonals are an **o** and an **s** for frame 2. The length of these diagonals equals the time τ elapsed on a clock that is at rest in frame 2, i.e., a clock that glides through space at constant velocity from A to B, reuniting with the planet earth when its orbit brings it to B. The area of the gray regions can be interpreted as half the square of this gliding-clock time, which is called the proper time. "Proper" is used here in the somewhat archaic sense of "own" or "self," as in "The Vatican does not lie within Italy proper." Proper time, which we notate τ , can only be defined for timelike world-lines, since a lightlike or spacelike world-line isn't possible for a material clock.

In terms of (Minkowski) coordinates, suppose that events A and B are separated by a distance x and a time t . Then in general $t^2 - x^2$ gives the square of the gliding-clock time. Proof: Because of the way that area scales with a rescaling of the coordinates, the expression must have the form $(\dots)t^2 + (\dots)tx + (\dots)x^2$, where each (\dots) represents a unitless constant. The tx coefficient must be zero by the isotropy of space. The t^2 coefficient must equal 1 in order to give the right answer in the case of $x = 0$, where the coordinates are those of an observer at rest relative to the clock. Since the area vanishes for $x = t$, the x^2 coefficient must equal -1 .

When $|x|$ is greater than $|t|$, events A and B are so far apart in space and so close together in time that it would be impossible to have a cause and effect relationship between them, since $c = 1$ is the maximum speed of cause and effect. In this situation $t^2 - x^2$ is negative and cannot be interpreted as a clock time, but it can be interpreted as minus the square of the distance between A and B, as measured in a frame of reference in which A and B are simultaneous.

Generalizing to $3+1$ dimensions and to any vector **v**, not just

a displacement in spacetime, we have a measurement of the vector defined by

$$v_t^2 - v_x^2 - v_y^2 - v_z^2.$$

In the special case where \mathbf{v} is a spacetime displacement, this can be referred to as the spacetime interval. Except for the signs, this looks very much like the Pythagorean theorem, which is a special case of the vector dot product. We therefore define a function g called the *metric*,

$$g(\mathbf{u}, \mathbf{v}) = u_t v_t - u_x v_x - u_y v_y - u_z v_z.$$

Because of the analogy with the Euclidean dot product, we often use the notation $\mathbf{u} \cdot \mathbf{v}$ for this quantity, and we sometimes call it the inner product. The metric is the central object of relativity. In general relativity, which describes gravity as a curvature of spacetime, the coefficients occurring on the right-hand side are no longer ± 1 , but must vary from point to point. Even in special relativity, where the coefficients can be made constant, the definition of g is arbitrary up to a nonzero multiplicative constant, and in particular many authors define g as the negative of our definition. The sign convention we use is the most common one in particle physics, while the opposite is more common in classical relativity. The set of signs, $+- --$ or $-+++$, is called the signature of the metric.

In subsection 1.1.3 we developed the idea of orthogonality of spacetime vectors, with the physical interpretation that if an observer moves along a vector \mathbf{o} , a vector \mathbf{s} that is orthogonal to \mathbf{o} is a vector of simultaneity. This corresponds to the vanishing of the inner product, $\mathbf{o} \cdot \mathbf{s} = 0$, and is only imperfectly analogous to the idea that Euclidean vectors are perpendicular if their dot product is zero. In particular, a nonzero Euclidean vector is never perpendicular to itself, but for any lightlike vector \mathbf{v} we have $\mathbf{v} \cdot \mathbf{v} = 0$. The metric doesn't give us a measure of the length of lightlike vectors. Physically, neither a ruler nor a clock can measure such a vector.

The metric in SI units

Example 5

Units with $c = 1$ are known as natural units. (They are natural to relativity in the same sense that units with $\hbar = 1$ are natural to quantum mechanics.) Any equation expressed in natural units can be reexpressed in SI units by the simple expedient of inserting factors of c wherever they are needed in order to get units that make sense. The result for the metric could be

$$g(\mathbf{u}, \mathbf{v}) = c^2 u_t v_t - u_x v_x - u_y v_y - u_z v_z$$

or

$$g(\mathbf{u}, \mathbf{v}) = u_t v_t - (u_x v_x - u_y v_y - u_z v_z)/c^2.$$

It doesn't matter which we pick, since the metric is arbitrary up to a constant factor. The former expression gives a result in meters, the latter seconds.

Orthogonal light rays?

Example 6

▷ On a spacetime diagram in 1+1 dimensions, we represent the light cone with the two lines $x = \pm t$, drawn at an angle of 90 degrees relative to one another. Are these lines orthogonal?

▷ No. For example, if $\mathbf{u} = (1, 1)$ and $\mathbf{v} = (1, -1)$, then $\mathbf{u} \cdot \mathbf{v}$ is 2, not zero.

Pioneer 10

Example 7

▷ The Pioneer 10 space probe was launched in 1972, and in 1973 was the first craft to fly by the planet Jupiter. It crossed the orbit of the planet Neptune in 1983, after which telemetry data were received until 2002. The following table gives the spacecraft's position relative to the sun at exactly midnight on January 1, 1983 and January 1, 1995. The 1983 date is taken to be $t = 0$.

t (s)	x	y	z
0	1.784×10^{12} m	3.951×10^{12} m	0.237×10^{12} m
3.7869120000×10^8 s	2.420×10^{12} m	8.827×10^{12} m	0.488×10^{12} m

Compare the time elapsed on the spacecraft to the time in a frame of reference tied to the sun.

▷ We can convert these data into natural units, with the distance unit being the second (i.e., a light-second, the distance light travels in one second) and the time unit being seconds. Converting and carrying out this subtraction, we have:

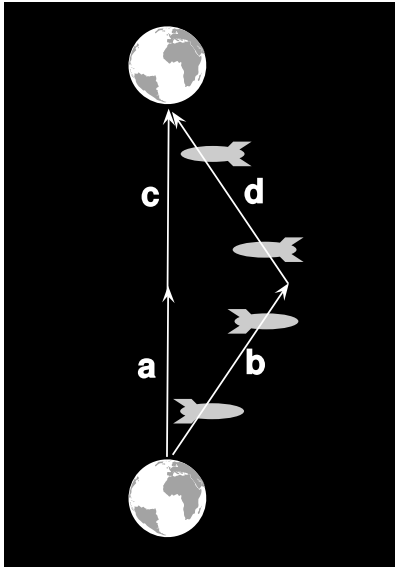
Δt (s)	Δx	Δy	Δz
3.7869120000×10^8 s	0.2121×10^4 s	1.626×10^4 s	0.084×10^4 s

Comparing the exponents of the temporal and spatial numbers, we can see that the spacecraft was moving at a velocity on the order of 10^{-4} of the speed of light, so relativistic effects should be small but not completely negligible.

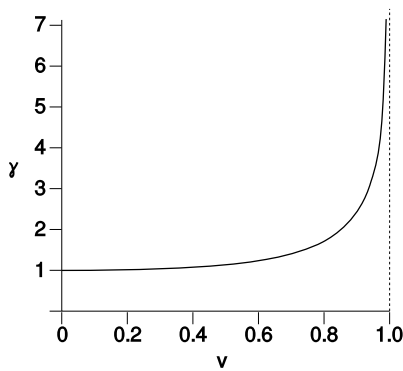
Since the interval is timelike, we can take its square root and interpret it as the time elapsed on the spacecraft. The result is $\tau = 3.786911996 \times 10^8$ s. This is 0.4 s less than the time elapsed in the sun's frame of reference.

1.3.3 The gamma factor

Figure s is the relativistic version of example 1 on p. 14. We intend to analyze it using the metric, and since the metric gives the same result in any frame, we have chosen for convenience to represent it in the frame in which the earth is at rest. We have $\mathbf{a} = (t, 0)$ and $\mathbf{b} = (t, vx)$, where v is the velocity of the spaceship relative to the earth. Application of the metric gives proper time t for the earthbound twin and $t\sqrt{1-v^2}$ for the traveling twin. The same results apply for \mathbf{c} and \mathbf{d} . The result is that the earthbound twin experiences a time that is greater by a factor γ (Greek letter gamma) defined as $\gamma = 1/\sqrt{1-v^2}$. If v is close to c , γ can be large, and we find that when the astronaut twin returns home, still youth-



s / The twin “paradox.”



t / A graph of γ as a function of v .

ful, the earthbound twin can be old and gray. This was at one time referred to as the twin paradox, and it was considered paradoxical either because it seemed to defy common sense or because the traveling twin could argue that she was the one at rest while the earth was moving. The violation of common sense is in fact what was observed in the Hafele-Keating experiment, and the latter argument is fallacious for the same reasons as in the Galilean version given in example 1.

We have in general the following interpretation:

Time dilation

A clock runs fastest in the frame of reference of an observer who is at rest relative to the clock. An observer in motion relative to the clock at speed v perceives the clock as running more slowly by a factor of γ .

Although this is phrased in terms of clocks, we interpret it as telling us something about time itself. The attitude is that we should define a concept in terms of the operations required in order to measure it: time is defined as what a clock measures. This philosophy, which has been immensely influential among physicists, is called operationalism and was developed by P.W. Bridgman in the 1920's. Our operational definition of time works because the rates of *all* physical processes are affected equally by time dilation.⁶ By the time the twins in figure s are reunited, not only has the traveling twin heard fewer ticks from her antique mechanical pocket watch, but she has also had fewer heartbeats, and the ship's atomic clock agrees with her watch to within the precision of the watch.

self-check A

What is γ when $v = 0$? What does this mean? Express the equation for γ in SI units. ▷ Answer, p. ??

Time dilation is symmetrical in the sense that it treats all frames of reference democratically. If observers A and B aren't at rest relative to each other, then A says B's time runs slow, but B says A is the slow one. In figure s, the laws of physics make no distinction between the frames of reference that coincide with vectors **a** and **b**; as in the corresponding Galilean case of example 1 on p. 14, the asymmetry comes about because **a** and **c** are parallel, but **b** and **d** are not.

⁶For more on this topic, see section 6.1.

As shown in example 8 below, consistency demands that in addition to the effect on time we have a similar effect on distances:

Length contraction

A meter-stick appears longest to an observer who is at rest relative to it. An observer moving relative to the meter-stick at v observes the stick to be shortened by a factor of γ .

The visualization of length contraction in terms of spacetime diagrams is presented in figure z/2 on p. 30. Our present discussion is limited to 1+1 dimensions, but in 3+1, only the length along the line of motion is contracted (ch. 2, problem 2, p. 51).

An interstellar road trip

Example 8

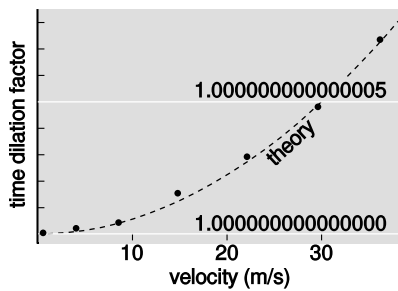
Alice stays on earth while her twin Betty heads off in a spaceship for Tau Ceti, a nearby star. Tau Ceti is 12 light-years away, so even though Betty travels at 87% of the speed of light, it will take her a long time to get there: 14 years, according to Alice.



u / Example 8.

Betty experiences time dilation. At this speed, her γ is 2.0, so that the voyage will only seem to her to last 7 years. But there is perfect symmetry between Alice's and Betty's frames of reference, so Betty agrees with Alice on their relative speed. (For more detail on this point, see example 11, p. 31.) Betty sees herself as being at rest, while the sun and Tau Ceti both move backward at 87% of the speed of light. How, then, can she observe Tau Ceti to get to her in only 7 years, when it should take 14 years to travel 12 light-years at this speed?

We need to take into account length contraction. Betty sees the distance between the sun and Tau Ceti to be shrunk by a factor of 2. The same thing occurs for Alice, who observes Betty and her spaceship to be foreshortened.



v / Time dilation measured with an atomic clock at low speeds. The theoretical curve, shown with a dashed line, is calculated from $\gamma = 1/\sqrt{1-v^2/c^2}$; at these small velocities, the approximation $\gamma \approx 1 + v^2/2c^2$ is excellent, and the graph is indistinguishable from a parabola. This graph corresponds to an extreme close-up view of the lower left corner of figure t. The error bars on the experimental points are about the same size as the dots.

A moving atomic clock

Example 9

Expanding γ in a Taylor series, we find $\gamma \approx 1 + v^2/2c^2$, so that when v is small, relativistic effects are approximately proportional to v^2 , so it is very difficult to observe them at low speeds. This was the reason that the Hafele-Keating experiment was done aboard passenger jets, which fly at high speeds. Jets, however, fly at high altitude, and this brings in a second time dilation effect, a general-relativistic one due to gravity. The main purpose of the experiment was actually to test this effect.

It was not until four decades after Hafele and Keating that anyone did a conceptually simple atomic clock experiment in which the only effect was motion, not gravity. In 2010, however, Chou *et al.*⁷ succeeded in building an atomic clock accurate enough to detect time dilation at speeds as low as 10 m/s. Figure v shows their results. Since it was not practical to move the entire clock, the experimenters only moved the aluminum atoms inside the clock that actually made it “tick.”

Large time dilation

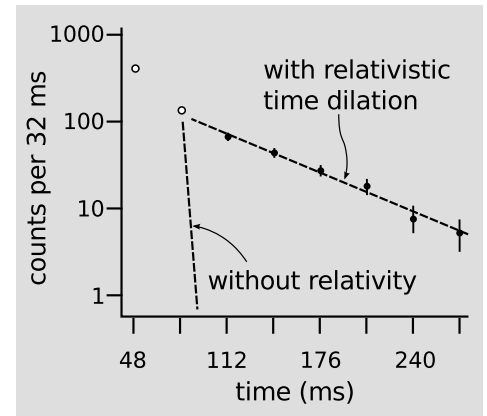
Example 10

The time dilation effects described in example 9 were very small. If we want to see a large time dilation effect, we can’t do it with something the size of the atomic clocks they used; the kinetic energy would be greater than the total megatonnage of all the world’s nuclear arsenals. We can, however, accelerate subatomic particles to speeds at which γ is large. For experimental particle physicists, relativity is something you do all day before heading home and stopping off at the store for milk. An early, low-precision experiment of this kind was performed by Rossi and Hall in 1941, using naturally occurring cosmic rays. Figure w shows a 1974 experiment⁸ of a similar type which verified the time dilation predicted by relativity to a precision of about one part per thousand.

Particles called muons (named after the Greek letter μ , “myoo”) were produced by an accelerator at CERN, near Geneva. A muon is essentially a heavier version of the electron. Muons undergo radioactive decay, lasting an average of only $2.197 \mu\text{s}$ before they evaporate into an electron and two neutrinos. The 1974 experiment was actually built in order to measure the magnetic properties of muons, but it produced a high-precision test of time dilation as a byproduct. Because muons have the same electric charge as electrons, they can be trapped using magnetic fields. Muons were injected into the ring shown in figure w, circling around it until they underwent radioactive decay. At the speed at which these muons were traveling, they had $\gamma = 29.33$, so on the average they lasted 29.33 times longer than the normal lifetime. In other words,

⁷Science 329 (2010) 1630

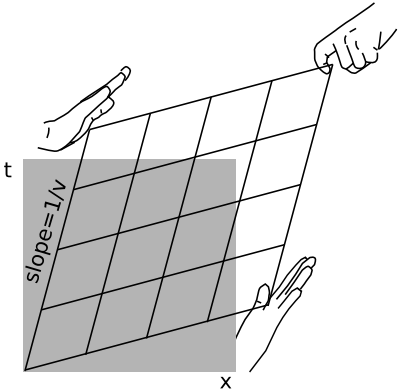
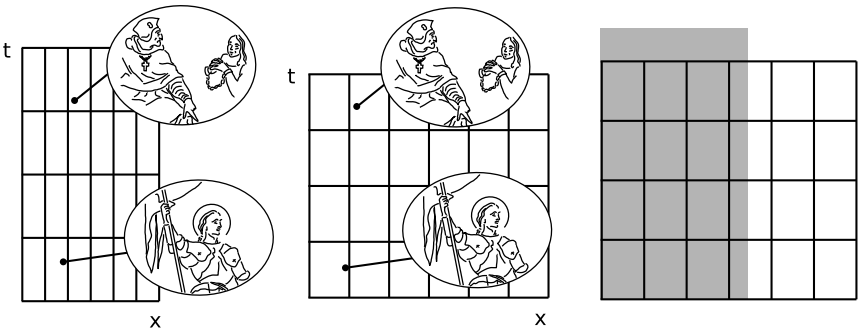
⁸Bailey *et al.*, Nucl. Phys. B150(1979) 1



w / *Left:* Apparatus used for the test of relativistic time dilation described in example 10. The prominent black and white blocks are large magnets surrounding a circular pipe with a vacuum inside. (c) 1974 by CERN. *Right:* Muons accelerated to nearly c undergo radioactive decay much more slowly than they would according to an observer at rest with respect to the muons. The first two data-points (unfilled circles) were subject to large systematic errors.

they were like tiny alarm clocks that self-destructed at a randomly selected time. The graph shows the number of radioactive decays counted, as a function of the time elapsed after a given stream of muons was injected into the storage ring. The two dashed lines show the rates of decay predicted with and without relativity. The relativistic line is the one that agrees with experiment.

x / Two events are given as points on a graph of position versus time. Joan of Arc helps to restore Charles VII to the throne. At a later time and a different position, Joan of Arc is sentenced to death.



y / The Lorentz transformation.

z / 1. The clock is at rest in the original frame of reference, and it measures a time interval t . In the new frame of reference, the time interval is greater by a factor of γ . 2. The ruler is moving in the first frame, represented by a square, but at rest in the second one, shown as a parallelogram. Each picture of the ruler is a snapshot taken at a certain moment as judged according to the second frame's notion of simultaneity. An observer in first frame judges the ruler's length instead according to that frame's definition of simultaneity, i.e., using points that are lined up horizontally on the graph. The ruler appears shorter in the frame in which it is moving.

1.4 The Lorentz transformation

Philosophically, coordinates are unnecessary, but they are convenient. They are arbitrary, so we can change from one set to another. For example, we can change the units used to measure time and position, as in the first and second panels of figure x. Nothing changes about the underlying events; only the labels are different. The third panel shows a convenient convention we will use to depict such changes visually. The gray rectangle represents the original grid from the first panel, while the grid of black lines represents the new version from the second panel. Omitting the grid from the gray rectangle makes the diagram easier to decode visually.

In special relativity it is of interest to convert between the Minkowski coordinates of observers who are in motion relative to one another. The result, shown in figure y, is a kind of stretching and smooshing of the diagonals. Since the area is invariant, one diagonal grows by the same factor by which the other shrinks. This change of coordinates is called the Lorentz transformation.

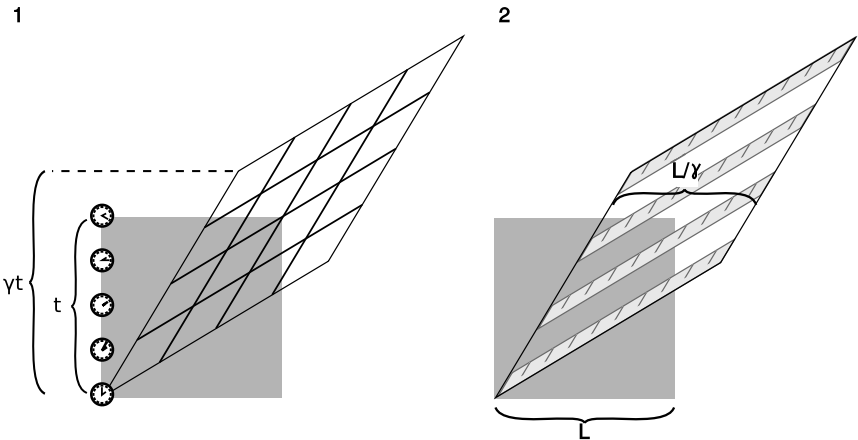


Figure z shows how time dilation and length contraction come about in this picture. It should be emphasized here that the Lorentz transformation includes more effects than just length contraction

and time dilation. Many beginners at relativity get confused and come to erroneous conclusions by trying to reduce everything to a matter of inserting factors of γ in various equations. If the Lorentz transformation amounted to nothing more than length contraction and time dilation, it would be merely a change of units like the one shown in figure x.

The Lorentz transformation can be notated algebraically:

$$\begin{aligned} t' &= \gamma t - v\gamma x \\ x' &= -v\gamma t + \gamma x \end{aligned} \quad (1)$$

The fact that this is the correct relativistic transformation can be verified by noting that (1) the speed-of-light lines $x = \pm t$ are preserved, and (2) the determinant equals 1, so that areas are preserved. Alternatively, it is sufficient to check the invariance of the spacetime interval under this transformation.

Equations (1) treat space and time in a perfectly symmetric way, but this should not be taken as implying that special relativity perfectly embodies such a symmetry. For example, I can easily revisit a place that I've been to before, but I can't go back in time. And of course we have three dimensions of space; our use of 1+1 dimensions rather than 3+1 is just a matter of convenience for the moment. Note also that there is no exact analogy between figure z/1, where the clock is a pointlike object tracing a line through spacetime, and z/2, where the ruler is an extended body that sweeps out a parallel-sided ribbon.

Observers agree on their relative speeds

Example 11

Observer A says observer B is moving away from her at velocity v ; is it true, as in Galilean relativity, that B measures the same speed for A? Yes, it is true, but not completely obvious. One way to verify this fact is to check that Lorentz transformations with velocities v and $-v$ are inverses. A more physically transparent justification is shown in figure aa. In aa/1, A determines B's velocity relative to her by sending out two round-trip signals at the speed of light, and measuring the difference between the two round-trip times. Because space is the same in all directions,⁹ the experimental data are exactly the same when B carries out the measurement, aa/2, and therefore B infers the same speed.

Motion in the opposite direction

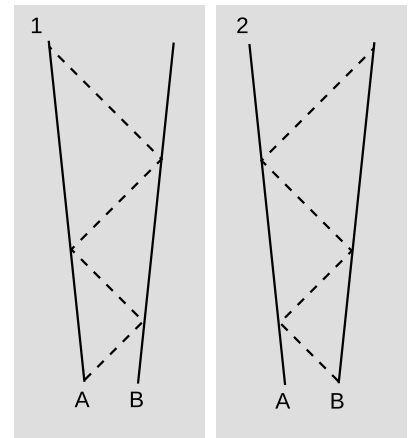
Example 12

Figure ab shows the case where the observer whose frame is represented by the grid is moving to the left relative to the one whose frame is represented by the gray square.

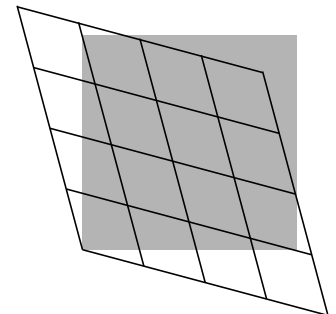
Other quadrants

Example 13

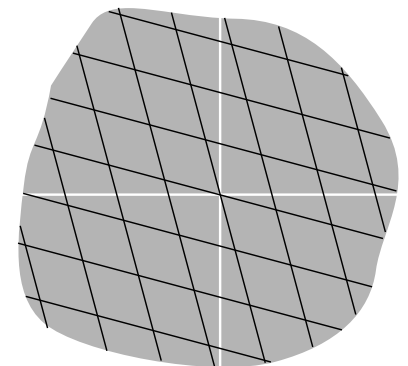
So far I've been arbitrarily choosing to draw only the first quadrant of each coordinate system. Figure ac shows a region that includes all four quadrants.



aa / Example 11.



ab / Example 12.



ac / Example 13.

⁹This is discussed in more detail on p. 46

A numerical example of invariance

Example 14

Figure 14.1 shows two frames of reference in motion relative to one another at $v = 3/5$. (For this velocity, the stretching and squishing of the main diagonals are both by a factor of 2.) Events are marked at coordinates that in the frame represented by the square are

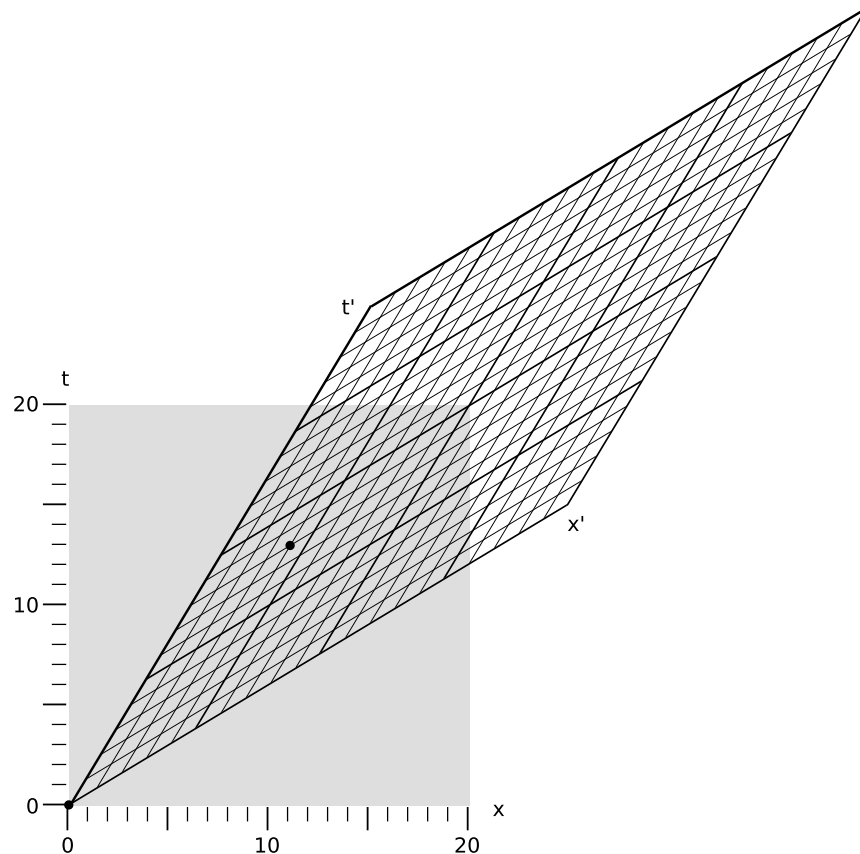
$$(t, x) = (0, 0) \quad \text{and} \\ (t, x) = (13, 11).$$

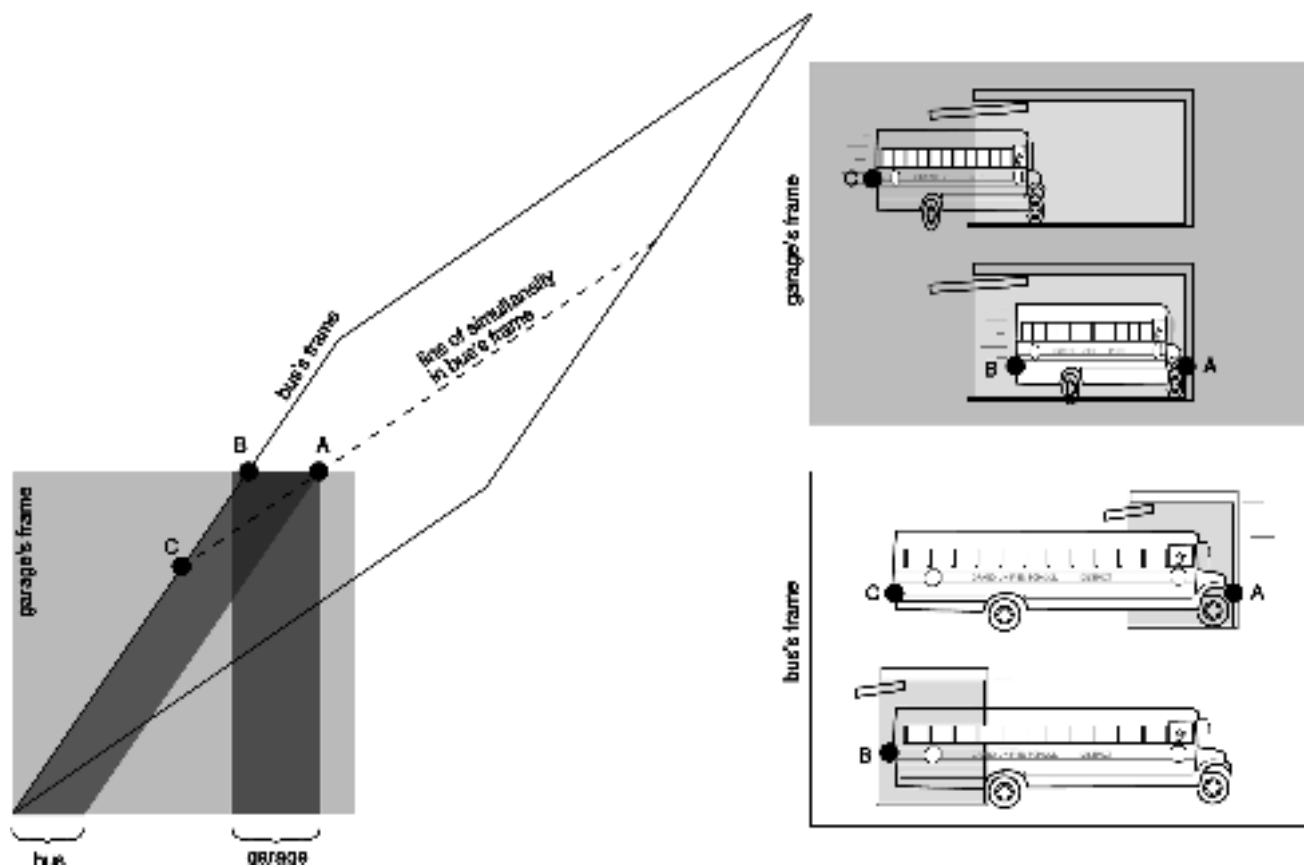
The interval between these events is $13^2 - 11^2 = 48$. In the frame represented by the parallelogram, the same two events lie at coordinates

$$(t', x') = (0, 0) \quad \text{and} \\ (t', x') = (8, 4).$$

Calculating the interval using these values, the result is $8^2 - 4^2 = 48$, which comes out the same as in the other frame.

ad / Example 14.





ae / Example 15: In the garage's frame of reference, the bus is moving, and can fit in the garage due to its length contraction. In the bus's frame of reference, the garage is moving, and can't hold the bus due to *its* length contraction.

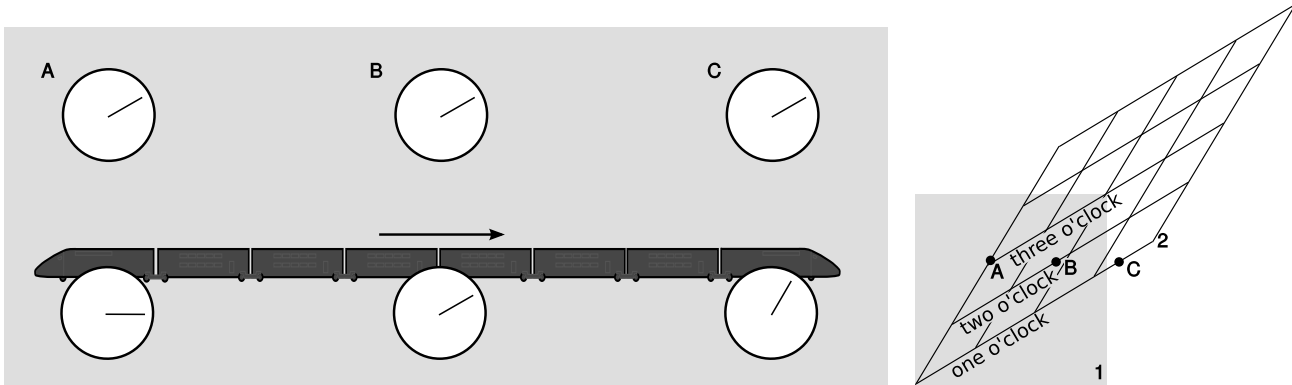
The garage paradox

Example 15

One of the most famous of all the so-called relativity paradoxes has to do with our incorrect feeling that simultaneity is well defined. The idea is that one could take a schoolbus and drive it at relativistic speeds into a garage of ordinary size, in which it normally would not fit. Because of the length contraction, the bus would supposedly fit in the garage. The driver, however, will perceive the *garage* as being contracted and thus even less able to contain the bus.

The paradox is resolved when we recognize that the concept of fitting the bus in the garage “all at once” contains a hidden assumption, the assumption that it makes sense to ask whether the front and back of the bus can *simultaneously* be in the garage. Observers in different frames of reference moving at high relative speeds do not necessarily agree on whether things happen simultaneously. As shown in figure ae, the person in the garage's frame can shut the door at an instant B he perceives to be si-

multaneous with the front bumper's arrival A at the back wall of the garage, but the driver would not agree about the simultaneity of these two events, and would perceive the door as having shut long after she plowed through the back wall.



af / Example 16.

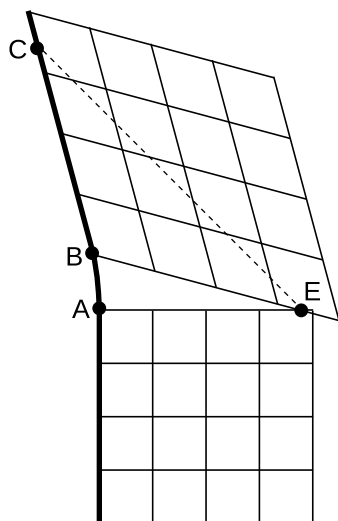
Shifting clocks

Example 16

The top row of clocks in the figure are located in three different places. They have been synchronized in the frame of reference of the earth, represented by the paper. This synchronization is carried out by exchanging light signals (Einstein synchronization), as in example 4 on p. 18. For example, if the front and back clocks both send out flashes of light when they think it's 2 o'clock, the one in the middle will receive them both at the same time. Event A is the one at which the back clock A reads 2 o'clock, etc.

The bottom row of clocks are aboard the train, and have been synchronized in a similar way. For the reasons discussed in example 4, their synchronization differs from that of the earth-based clocks. By referring to the diagram of the Lorentz transformation shown on the right, we see that in the frame of the train, 2, C happens first, then B, then A.

This is an example of the interpretation of the term $t' = \dots - v\gamma x$ in the Lorentz transformation (eq. (1), p. 31). Because the events occur at different x 's, each is shifted in time relative to the next, according to clocks synchronized in frame 2 (t' , the train).



ag / Example 17.

Deja vu?

Example 17

The grids we've been drawing are mere *conventions*, as elaborate and arbitrary as dress in the court of Louis XIV. They come from a surveying process, which may need to be planned in advance and whose results we might not be able to see until later.

The dark line in figure ag is the world-line of an observer O, who moves inertially for a while, accelerates to the left, and then

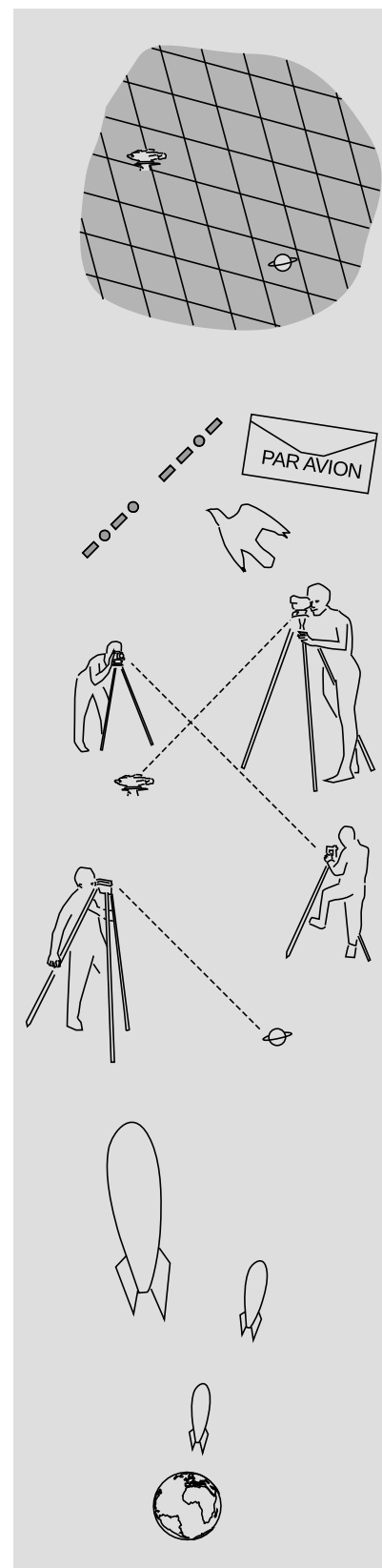
moves inertially again. On the right are pasted two coordinate grids adapted to the two inertial segments. At event E, Bush steals the 2000 election, and this is depicted as being simultaneous with both event A and event B. Does poor O see it happen twice? No, even if the bad news is transmitted by a signal moving at the speed of light (dashed line), O receives it only once, at event C.

The only problem here is a poor choice of labels, which causes E to have more than one label. Something similar happens in a constant-acceleration frame, section 7.1, p. 143. Cf. also p. 73.

Many mistakes by beginners at relativity revolve around a set of unexamined preconceptions about what it means to observe things. One imagines that effects such as length contraction and time dilation are what an observer actually *sees*, and perhaps that this process of seeing is instantaneous. Or one thinks of Minkowski coordinates as if they were the result of a simple and automatic process of perception by an observer. This is the kind of thinking that will lead one to believe that example 17 is crazy or paradoxical.

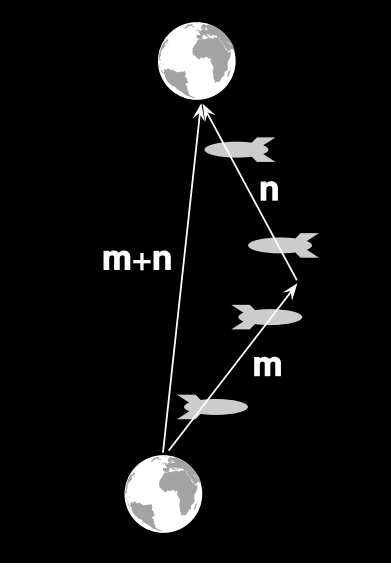
As another example, it should not be imagined that the length contraction of a stick by $1/\gamma$ is what an observer actually *sees* when looking at the stick. Optical observations are influenced, for example, by the unequal times taken for light to propagate from the ends of the stick to the eye. A simulation of this type of effect is drawn in example 7 on p. 135.

Length contraction, time dilation, the observer-dependence of simultaneity, and Minkowski coordinates are all sophisticated results of a laborious process of collecting and analyzing data obtained by techniques such as Einstein synchronization, which require actions such as consulting atomic clocks or exchanging signals between different points at the speed of light. Figure ah outlines such a process in a cartoonish way. A fleet of rocket ships, carrying surveyors, is sent out from Earth and dispersed throughout a vast region of space. Surveyors look through their theodolites at images, which are formed by light rays (dashed lines) that have arrived after traveling at the finite speed c . Such light rays carry old, stale information about various events. A nuclear war has broken out. Rock and roll music has arrived on Saturn. The resulting data are then transmitted by various means (passenger pigeon, Morse-coded radio, paper mail) and consolidated at the surveying office. At the office, workers at a long row of desks crunch the numbers and produce a chart of Minkowski coordinates with the events marked in.



ah / Minkowski coordinates are the result of a complicated process of surveying.

1.5 ★ Triangle and Cauchy-Schwarz inequalities



ai / The twin paradox, interpreted as a triangle inequality.

In Euclidean geometry, we have the intuitively obvious fact that any side of a triangle is no greater than the sum of the other two sides. This can be written in terms of vectors as $|\mathbf{m} + \mathbf{n}| \leq |\mathbf{m}| + |\mathbf{n}|$. Closely related to it is the inequality $|\mathbf{m} \cdot \mathbf{n}| \leq |\mathbf{m}| |\mathbf{n}|$, known as the Cauchy-Schwarz inequality, which can be seen because $\mathbf{m} \cdot \mathbf{n} = |\mathbf{m}| |\mathbf{n}| \cos \theta$, where θ is the angle between the two vectors.

Any proof of these facts ultimately depends on the assumption that the metric has the Euclidean signature $+++$ (or on equivalent assumptions such as Euclid's axioms). Figure ai shows that on physical grounds, we do not expect the inequalities to hold for Minkowski vectors in their unmodified Euclidean forms. The quantity $|\mathbf{m} + \mathbf{n}|$ represents the proper time of the spaceship that moved inertially along with the earth, while $|\mathbf{m}| + |\mathbf{n}|$ is the *greater* proper time of the traveling spaceship.

On the other hand, Minkowski space has copies of Euclidean space built in. For example, we know that all the familiar Euclidean facts must hold in any plane of simultaneity defined by a particular observer at a given moment in time, since the restriction of the metric to that plane has signature $---$, and the distinction between this and the $+++$ signature is an arbitrary notational convention.

Summarizing these observations, we expect that the relativistic version of the triangle and Cauchy-Schwarz inequalities will be split into cases, some of which are the same as the Euclidean case and some of them different.

Some notational issues may be confusing in the following discussion. We let \mathbf{a}^2 mean $\mathbf{a} \cdot \mathbf{a}$, which may not be positive, while $|\mathbf{a}|$ indicates the positive real number $\sqrt{|\mathbf{a} \cdot \mathbf{a}|}$. I will try to specifically point out any equations that are only true for $+---$ signature and not for $-+++$, and express important final results in a way that doesn't depend on this choice.

1.5.1 Two timelike vectors

A simple and important case is the one in which both \mathbf{m} and \mathbf{n} trace possible world-lines of material objects, as in figure ai. That is, they must both be timelike vectors. To see what form of the Cauchy-Schwarz inequality should hold, we break the vector \mathbf{n} down into two parts, $\mathbf{n} = \mathbf{n}_{\parallel} + \mathbf{n}_{\perp}$, where \mathbf{n}_{\parallel} is parallel to \mathbf{m} and \mathbf{n}_{\perp} perpendicular. We then have $|\mathbf{m} \cdot \mathbf{n}| = |\mathbf{m} \cdot \mathbf{n}_{\parallel}| = |\mathbf{m}| |\mathbf{n}_{\parallel}|$. But $\mathbf{n}^2 = (\mathbf{n}_{\parallel} + \mathbf{n}_{\perp})^2 = \mathbf{n}_{\parallel}^2 + 2\mathbf{n}_{\parallel} \cdot \mathbf{n}_{\perp} + \mathbf{n}_{\perp}^2 = \mathbf{n}_{\parallel}^2 + \mathbf{n}_{\perp}^2$, and since \mathbf{n}_{\parallel} is timelike and \mathbf{n}_{\perp} spacelike, we have (in the $+---$ signature) $\mathbf{n}_{\parallel}^2 > 0$ and $\mathbf{n}_{\perp}^2 < 0$. Therefore, regardless of signature, $|\mathbf{n}| \leq |\mathbf{n}_{\parallel}|$, and we have the reversed Cauchy-Schwarz inequality

$$|\mathbf{m} \cdot \mathbf{n}| \geq |\mathbf{m}| |\mathbf{n}| \quad (\text{valid for either } +--- \text{ or } -+++).$$

A useful way of interpreting the reversal compared to the Euclidean case is that if the vectors happen to be normalized such that $|\mathbf{m}| = |\mathbf{n}| = 1$, then $\mathbf{m} \cdot \mathbf{n} = \gamma$, where γ is the Lorentz factor for an observer whose world-line is parallel to \mathbf{m} with respect to a world-line parallel to \mathbf{n} . The difference from the Euclidean behavior can then be understood as arising from the fact that whereas $|\cos \theta| \leq 1$, we always have $\gamma \geq 1$.

Given the physical motivation presented so far, it would have been natural to take both \mathbf{m} and \mathbf{n} to lie in the future rather than the past light cone, but we have not yet assumed that this was the case, and the reversed Cauchy-Schwarz inequality holds independently of such an assumption. (See problem 16 for an alternative way of seeing this.) In order to discuss the related triangle inequality, however, we will need to assume that both vectors are future-directed. Physically, this is necessary in order to give the interpretation shown in figure ai, from which we have already inferred that the triangle inequality must be reversed. To verify this mathematically, we can compute the difference $(\mathbf{m} + \mathbf{n})^2 - (|\mathbf{m}| + |\mathbf{n}|)^2$ (problem 17).

An application to collisions is given in section 4.3.2, p. 89.

1.5.2 Two spacelike vectors not spanning the light cone

Now suppose that \mathbf{m} and \mathbf{n} are both spacelike, and the plane that they span does not include the light-cone. Operating within this plane, we never get any timelike or lightlike vectors, and therefore the non-Euclidean nature of the metric is never apparent to us. The geometry of this plane is therefore Euclidean, so in this case the ordinary Euclidean versions of the Cauchy-Schwarz and triangle inequalities must hold.

No relativity required

Example 18

Suppose that a certain observer establishes Minkowski coordinates, and consider the unit vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ lying along the x and y axes. The x - y plane that they span does not include the light cone. By plugging in to the Minkowski-coordinate form of the metric, we find that $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = 0$, as expected since the geometry of the x - y plane is Euclidean. This satisfies the ordinary form of the Cauchy-Schwarz inequality.

1.5.3 Two spacelike vectors spanning the light cone

Now consider the case, in Minkowski coordinates, where $\mathbf{m} = (0, 5, 0, 0)$ and $\mathbf{n} = (4, 5, 0, 0)$. These vectors span the t - x plane, whose geometry is not Euclidean, and they do not satisfy the Euclidean Cauchy-Schwarz inequality, since $\mathbf{m} \cdot \mathbf{n} = -25$, whereas $|\mathbf{m}| |\mathbf{n}| = 15$. Two vectors of this type will always satisfy the reversed version of the Cauchy-Schwarz inequality (problem 18). The converse holds in the sense that if two spacelike vectors satisfy the strict inequality $|\mathbf{m} \cdot \mathbf{n}| > |\mathbf{m}| |\mathbf{n}|$, then they span the light cone.

Problems

1 Astronauts in three different spaceships are communicating with each other. Those aboard ships A and B agree on the rate at which time is passing, but they disagree with the ones on ship C.

(a) Alice is aboard ship A. How does she describe the motion of her own ship, in its frame of reference?

(b) Describe the motion of the other two ships according to Alice.

(c) Give the description according to Betty, whose frame of reference is ship B.

(d) Do the same for Cathy, aboard ship C.

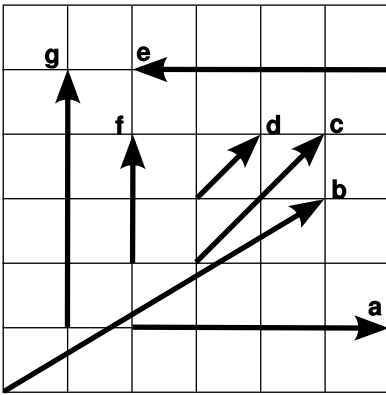
2 What happens in the equation for γ when you put in a negative number for v ? Explain what this means physically, and why it makes sense.

3 The Voyager 1 space probe, launched in 1977, is moving faster relative to the earth than any other human-made object, at 17,000 meters per second.

(a) Calculate the probe's γ .

(b) Over the course of one year on earth, slightly less than one year passes on the probe. How much less? (There are 31 million seconds in a year.) \checkmark

4 The earth is orbiting the sun, and therefore is contracted relativistically in the direction of its motion. Compute the amount by which its diameter shrinks in this direction. \checkmark



Problem 5.

5 The figure shows seven displacement vectors in spacetime. Which of these represent spacetime intervals that are equal to one another?

6 (a) In Euclidean geometry in three dimensions, suppose we have two vectors, \mathbf{a} and \mathbf{b} , which are unit vectors, i.e., $\mathbf{a} \cdot \mathbf{a} = 1$ and $\mathbf{b} \cdot \mathbf{b} = 1$. What is the range of possible values for the inner product $\mathbf{a} \cdot \mathbf{b}$?

(b) Repeat part a for two timelike, future-directed unit vectors in $3 + 1$ dimensions.

7 Expressed in natural units, the Lorentz transformation is

$$t' = \gamma t - v\gamma x$$

$$x' = -v\gamma t + \gamma x.$$

(a) Insert factors of c to make it valid in units where $c \neq 1$. (b) Show that in the limit $c \rightarrow \infty$, these have the right Galilean behavior.

8 This problem assumes you have some basic knowledge of quantum physics. One way of expressing the correspondence principle as applied to special relativity is that in the limit $c \rightarrow \infty$, all relativistic expressions have to go over to their Galilean counterparts. What would be the corresponding limit if we wanted to recover classical mechanics from quantum mechanics?

9 In $3 + 1$ dimensions, prove that if \mathbf{u} and \mathbf{v} are nonzero, future-lightlike, and not parallel to each other, then their sum is future-timelike.

10 Prove that if \mathbf{u} and \mathbf{v} are nonzero, lightlike, and orthogonal to each other, then they are parallel, i.e., $\mathbf{u} = c\mathbf{v}$ for some $c \neq 0$.

11 The speed at which a disturbance travels along a string under tension is given by $v = \sqrt{T/\mu}$, where μ is the mass per unit length, and T is the tension.

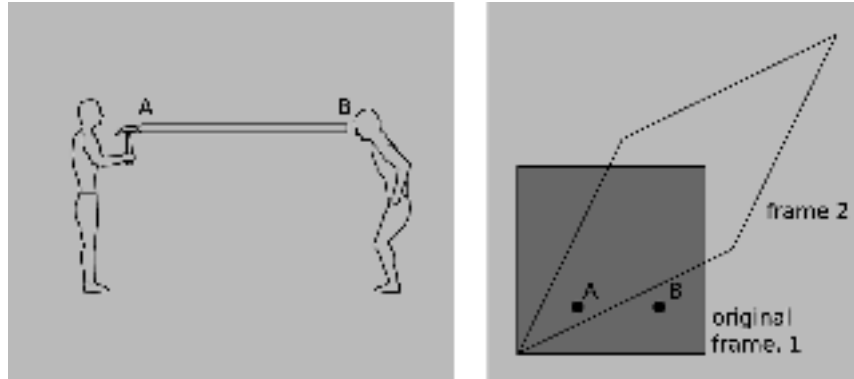
(a) Suppose a string has a density ρ , and a cross-sectional area A . Find an expression for the maximum tension that could possibly exist in the string without producing $v > c$, which is impossible according to relativity. Express your answer in terms of ρ , A , and c . The interpretation is that relativity puts a limit on how strong any material can be. \checkmark

(b) Every substance has a tensile strength, defined as the force per unit area required to break it by pulling it apart. The tensile strength is measured in units of N/m^2 , which is the same as the pascal (Pa), the mks unit of pressure. Make a numerical estimate of the maximum tensile strength allowed by relativity in the case where the rope is made out of ordinary matter, with a density on the same order of magnitude as that of water. (For comparison, kevlar has a tensile strength of about 4×10^9 Pa, and there is speculation that fibers made from carbon nanotubes could have values as high as 6×10^{10} Pa.) \checkmark

(c) A black hole is a star that has collapsed and become very dense, so that its gravity is too strong for anything ever to escape from it. For instance, the escape velocity from a black hole is greater than c , so a projectile can't be shot out of it. Many people, when they hear this description of a black hole in terms of an escape velocity greater than c , wonder why it still wouldn't be possible to extract an object from a black hole by other means than launching it out as a projectile. For example, suppose we lower an astronaut into a black hole on a rope, and then pull him back out again. Why might this not work?

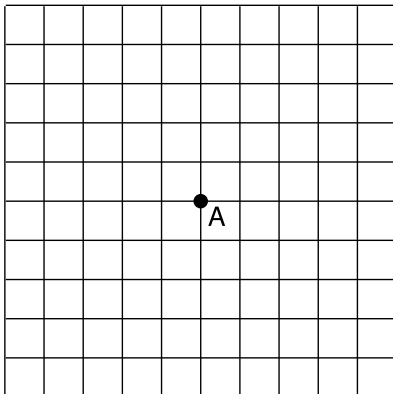
12 The rod in the figure is perfectly rigid. At event A, the hammer strikes one end of the rod. At event B, the other end moves. Since the rod is perfectly rigid, it can't compress, so A and B are simultaneous. In frame 2, B happens before A. Did the motion at the right end *cause* the person on the left to decide to pick up the hammer and use it?

Problem 12.



13 Use a spacetime diagram to resolve the following relativity paradox. Relativity says that in one frame of reference, event A could happen before event B, but in someone else's frame B would come before A. How can this be? Obviously the two people could meet up at A and talk as they cruised past each other. Wouldn't they have to agree on whether B had already happened?

14 The grid represents spacetime in a certain frame of reference. Event A is marked with a dot. Mark additional points satisfying the following criteria. (Pick points that lie at the intersections of the gridlines.)



Point B is at the same location as A in this frame of reference, and lies in its future.

C is also in point A's future, is not at the same location as A in this frame, but is in the same location as A according to some other frame of reference.

D is simultaneous with A in this frame of reference.

E is not simultaneous with A in this frame of reference, but is simultaneous with it according to some other frame.

F lies in A's past according to this frame of reference, but could not have caused A.

G lies in A's future according to this frame of reference, but is in its past according to some other frames.

H lies in A's future according to any frame of reference, not just this one.

Problem 14.

I is the departure of a spaceship, which arrives at A.

J could have caused A, but could not have been the departure of a spaceship like I that arrived later at A.

15 (a) Given an observer whose world-line is along a four-vector \mathbf{O} , suppose we want to determine whether some other four-vector \mathbf{P} is also a possible world-line of an observer. Show that knowledge of the signs of the inner products $\mathbf{O} \cdot \mathbf{P}$ and $\mathbf{P} \cdot \mathbf{P}$ is necessary and sufficient to determine this.

(b) Suppose that \mathbf{U} and \mathbf{V} are both observer-vectors. What would it mean physically to compute $\mathbf{U} + \mathbf{V}$?

(c) For vectors as described in part b, determine the signs of

$$(\mathbf{U} + \mathbf{V}) \cdot (\mathbf{U} + \mathbf{V})$$

and

$$(\mathbf{U} + \mathbf{V}) \cdot \mathbf{U}$$

by multiplying them out. Interpret the result physically.

16 In section 1.5.1, we proved the reversed Cauchy-Schwarz inequality for two timelike vectors, without any assumption as to whether they lay in the future or past light cones. But suppose that we had only established this fact for two vectors that were both future-directed. Show that the same inequality would then also have to hold regardless of whether one or both vectors was past-directed.

17 In the case of two future-directed timelike vectors, complete the proof of the reversed triangle inequality using the method suggested in section 1.5.1.

18 In section 1.5.3 we claimed that the reversed Cauchy-Schwarz inequality holds for two spacelike vectors that span the light cone. The purpose of this problem is to prove this fact using the following sketch of an argument provided by PhysicsForums user martinbn. Suppose that spacelike vectors \mathbf{m} and \mathbf{n} span the light cone, so that we can find some real number α such that $\mathbf{p} = \alpha\mathbf{m} + \mathbf{n}$ is lightlike. Compute \mathbf{p}^2 , and show that since α is real, the reversed Cauchy-Schwarz inequality holds.

19 A length-contracted object has length $L = L_0/\gamma$. Joe differentiates this with respect to time and finds $dL/dt = -L_0 v \gamma dv/dt$. He reasons that there is no upper limit on the magnitude of dv/dt , and therefore if $v \neq 0$ the quantity dL/dt can be arbitrarily large. This means that if an object accelerates away from an observer, its trailing edge can have $v > c$, which is supposed to be forbidden by relativity. OMG! Is Joe's reasoning correct?

Chapter 2

Foundations (optional)

In this optional chapter we more systematically examine the foundational assumptions of special relativity, which were appealed to casually in chapter 1. Most readers will want to skip this chapter and move on to ch. 3. The ordering of chapters 1 and 2 may seem backwards, but many of the issues to be raised here are very subtle and hard to appreciate without already understanding something about special relativity — in fact, Einstein and other relativists did not understand them properly until decades after the introduction of special relativity in 1905.

2.1 Causality

2.1.1 The arrow of time

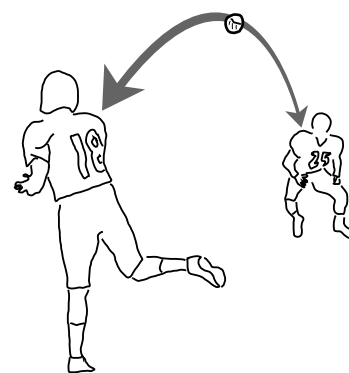
Our intuitive belief in cause-and-effect mechanisms is not supported in any clearcut way by the laws of physics as currently understood. For example, we feel that the past affects the future but not the other way around, but this feeling doesn't seem to translate into physical law. For example, Newton's laws are invariant under time reversal, figure a, as are Maxwell's equations. (The weak nuclear force is the only part of the standard model that violates time-reversal symmetry, and even it is invariant under the CPT transformation.)

There is an arrow of time provided by the second law of thermodynamics, and this arises ultimately from the fact that, for reasons unknown to us, the universe soon after the Big Bang was in a state of extremely low entropy.¹

2.1.2 Initial-value problems

So rather than depending on the arrow of time, we may be better off formulating a notion of causality based on existence and uniqueness of initial-value problems. In 1776, Laplace gave an influential early formulation of this idea in the context of Newtonian mechanics: "Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective posi-

¹One can find a vast amount of nonsense written about this, such as claims that the second law is derivable without reference to any cosmological context. For a careful treatment, see Callender, "Thermodynamic Asymmetry in Time," The Stanford Encyclopedia of Philosophy, plato.stanford.edu/archives/fall2011/entries/time-thermo.



a / Newton's laws do not distinguish past from future. The football could travel in either direction while obeying Newton's laws.

tions of the things which compose it . . . nothing would be uncertain, and the future as the past would be laid out before its eyes.” The reference to “one instant” is not compatible with special relativity, which has no frame-independent definition of simultaneity. We can, however, define initial conditions on some spacelike three-surface, i.e., a three-dimensional set of events that is smooth, has the topology of Euclidean space, and whose events are spacelike in relation to one another.

Unfortunately it is not obvious whether the classical laws of physics satisfy Laplace’s definition of causality. Two interesting and accessible papers that express a skeptical view on this issue are Norton, “Causation as Folk Science,” philsci-archive.pitt.edu/1214; and Echeverria *et al.*, “Billiard balls in wormhole spacetimes with closed timelike curves: Classical theory,” <http://resolver.caltech.edu/CaltechAUTHORS:ECHprd91>. The Norton paper in particular has generated a large literature at the interface between physics and philosophy, and one can find most of the relevant material online using the keywords “Norton’s dome.”

Nor does general relativity offer much support to the Laplacian version of causality. For example, general relativity says that given generic initial conditions, gravitational collapse leads to the formation of singularities, points where the structure of spacetime breaks down and various measurable quantities become infinite. Singularities typically violate causality, since the laws of physics can’t describe them. In a famous image, John Earman wrote that if we have a certain type of singularity (called a naked singularity), “all sorts of nasty things . . . emerge helter-skelter . . .,” including “TV sets showing Nixon’s ‘Checkers’ speech, green slime, Japanese horror movie monsters, etc.”

2.1.3 A modest definition of causality

Since there does not seem to be any reason to expect causality to hold in any grand sense, we will content ourselves here with a very modest and specialized definition, stated as a postulate, that works well enough for special relativity.

P1. *Causality.* There exist events 1 and 2 such that the displacement vector $\Delta \mathbf{r}_{12}$ is timelike in all frames.

This is sufficient to rule out the “rotational” version of the Lorentz transformation shown in figure j on p. 16. If P1 were violated, then we could never describe one event as causing another, since there would always be frames of reference in which the effect was observed as preceding the cause.

2.2 Flatness

2.2.1 Failure of parallelism

In postulate P1 we implicitly assumed that given two points, there was a certain vector connecting them. This is analogous to the Euclidean postulate that two points define a line.

For insight, let's think about how the Euclidean version of this assumption could fail. Euclidean geometry is only an approximate description of the earth's surface, for example, and this is why flat maps always entail distortions of the actual shapes. The distortions might be negligible on a map of Connecticut, but severe for a map of the whole world. That is, the globe is only locally Euclidean. On a spherical surface, the appropriate object to play the role of a "line" is a great circle, figure b. The lines of longitude are examples of great circles, and since these all coincide at the poles, we can see that two points do not determine a line in noneuclidean geometry.

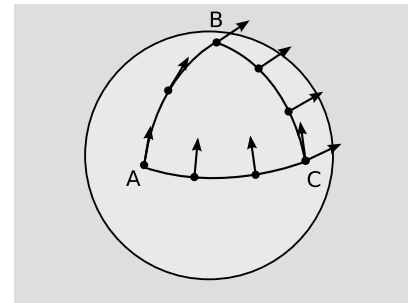
A two-dimensional bug living on the surface of a sphere would not be able to tell that the sphere was embedded in a third dimension, but it could still detect the curvature of the surface. It could tell that Euclid's postulates were false on large distance scales. A method that has a better analog in spacetime is shown in figure c: transporting a vector from one point to another depends on the path along which it was transported. This effect is our definition of curvature.

2.2.2 Parallel transport

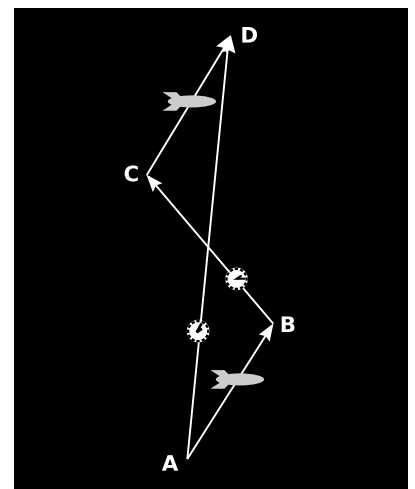
The particular type of transport that we have in mind here is called parallel transport. When I walk from the living room to the kitchen while carrying a mechanical gyroscope, I'm parallel-transporting the spacelike vector indicated by the direction of its axis. Figure d shows that parallel transport can also be defined for timelike vectors, and that parallel transport can be defined in spacetime using only inertial motion, clocks, and intersection of world-lines. Observers aboard the two spaceships use clocks in order to verify the parallelism of their world-lines (vectors AB and CD, which have equal lengths as measured by the proper time elapsed aboard the ships). Observer AB shoots clocks to observer CD, and the clocks are set up so that when they pass by one another, they automatically record one another's readings. The vectors are parallel if the record later reveals AD and BC intersected at their midpoints, as measured by the proper times recorded on the clocks.



b / An airplane flying from Mexico City to London follows the shortest path, which is a segment of a great circle. A path of extremal length between two points is called a geodesic.



c / Transporting the vector along path AC gives a different result than doing it along the path ABC.



d / Parallel transport.

2.2.3 Special relativity requires flat spacetime

Hidden in a number of spots in chapter 1 was the following assumption.

P2. *Flatness of spacetime.* Parallel-transporting a vector from one point to another gives a result that is independent of the path along which it was transported.

For example, when we established the form of the metric in section 1.3.2, we used the fact, proved on p. 49, that area is a scalar, but that proof depends on P2.

Property P2 is only approximately true, as shown explicitly by the Gravity Probe B satellite, launched in 2004. The probe carried four gyroscopes made of quartz, which were the most perfect spheres ever manufactured, varying from sphericity by no more than about 40 atoms. After one year and about 5000 orbits around the earth, the gyroscopes were found to have changed their orientations relative to the distant stars by about 3×10^{-6} radians (figure e). This is a violation of P2, but one that was very small and difficult to detect. The result was in good agreement with the predictions of general relativity, which describes gravity as a curvature of spacetime. The smallness of the effect tells us that the earth's gravitational field is not so large as to completely invalidate special relativity as a description of the nearby region of spacetime. One of the basic assumptions of general relativity is that in a small enough region of spacetime, it is always a good approximation to assume P2, so that general relativity is locally the same as special relativity. In the Gravity Probe B experiment, the effect was small and hard to detect, and this was the reason for letting the effect accumulate over a large number of orbits, spanning a large region of spacetime. Problem 5 on p. 52 investigates more quantitatively how the size of curvature effects varies with the size of the region.

2.3 Additional postulates

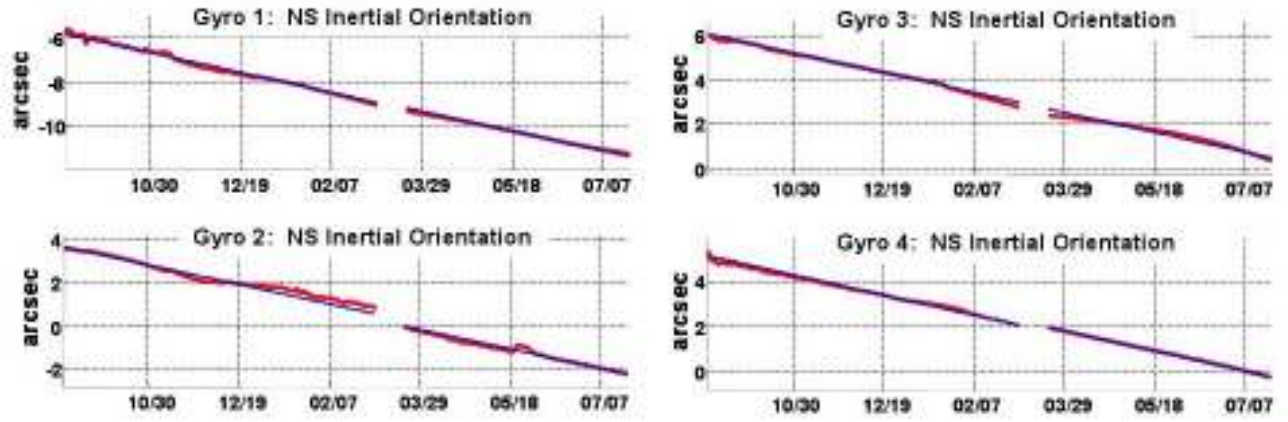
We make the following additional assumptions:

P3 *Spacetime is homogeneous and isotropic.* No time or place has special properties that make it distinguishable from other points, nor is one direction in space distinguishable from another.²

P4 *Inertial frames of reference exist.* These are frames in which particles move at constant velocity if not subject to any forces.³

²For the experimental evidence on isotropy, see http://www.edu-observatory.org/physics-faq/Relativity/SR/experiments.html#Tests_of_isotropy_of_space.

³Defining this no-force rule turns out to be tricky when it comes to gravity. As



e / Precession angle as a function of time as measured by the four gyroscopes aboard Gravity Probe B.

We can construct such a frame by using a particular particle, which is not subject to any forces, as a reference point. Inertial motion is modeled by vectors and parallelism.

P5 *Equivalence of inertial frames*: If a frame is in constant-velocity translational motion relative to an inertial frame, then it is also an inertial frame. No experiment can distinguish one preferred inertial frame from all the others.

P6 *Relativity of time*: There exist events 1 and 2 and frames of reference defined by observers \mathbf{o} and \mathbf{o}' such that $\mathbf{o} \perp \mathbf{r}_{12}$ is true but $\mathbf{o}' \perp \mathbf{r}_{12}$ is false, where the notation $\mathbf{o} \perp \mathbf{r}$ means that observer \mathbf{o} finds \mathbf{r} to be a vector of simultaneity according to some convenient criterion such as Einstein synchronization.⁴

Postulates P3 and P5 describe symmetries of spacetime, while P6 differentiates the spacetime of special relativity from Galilean spacetime; the symmetry described by these three postulates is referred to as Lorentz invariance, and all known physical laws have this symmetry. Postulate P4 defines what we have meant when we referred to the parallelism of vectors in spacetime (e.g., in figure s on p. 26). Postulates P1-P6 were all the assumptions that were needed in order to arrive at the picture of spacetime described in ch. 1. This approach, based on symmetries, dates back to 1911.⁵

Surprisingly, it is possible for space or spacetime to obey our flatness postulate P2 while nevertheless having a nontrivial *topology*,

discussed in ch. 5, this apparently minor technicality turns out to have important consequences.

⁴example 4, p. 18

⁵W. v. Ignatowsky, Phys. Zeits. 11 (1911) 972. English translation at en.wikisource.org/wiki/Translation:Some_General_Remarks_on_the_Relativity_Principle

such as that of a cylinder or a Möbius strip (cf. problem 4, p. 51, and sec. 7.6.2, p. 154). Many authors prefer to explicitly rule out such possibilities as part of their definition of special relativity.

2.4 Other axiomatizations

2.4.1 Einstein's postulates

Einstein used a different axiomatization in his 1905 paper on special relativity:⁶

E1. *Principle of relativity:* The laws of electrodynamics and optics are valid for all frames of reference for which the equations of mechanics hold good.

E2. Light is always propagated in empty space with a definite velocity c which is independent of the state of motion of the emitting body.

These should be supplemented with our P2 and P3.

Einstein's approach has been slavishly followed in many later textbook presentations, even though the special role it assigns to light is not consistent with how modern physicists think about the fundamental structure of the laws of physics. (In 1905 there was no other phenomenon known to travel at c .) Einstein did not explicitly state anything like our P2 (flatness), since he had not yet developed the theory of general relativity or the idea of representing gravity in relativity as spacetime curvature. When he did publish the general theory, he described the distinction between special and general relativity as a generalization of the class of acceptable frames of reference to include accelerated as well as inertial frames. This description has not stood the test of time, and today relativists use flatness as the distinguishing criterion. In particular, it is not true, as one sometimes still hears claimed, that special relativity is incompatible with accelerated frames of reference.

2.4.2 Maximal time

Another approach, presented, e.g., by Laurent,⁷ combines our P2 with the following:

T1 *Metric:* An inner product exists. Proper time is measured by the square root of the inner product of a world-line with itself.

T2 *Maximum proper time:* Inertial motion gives a world-line along which the proper time is at a maximum with respect to small changes in the world-line. Inertial motion is modeled by vectors and parallelism, and this vector-space apparatus has the

⁶Paraphrased from the translation by W. Perrett and G.B. Jeffery.

⁷Bertel Laurent, Introduction to Spacetime: A First Course on Relativity

usual algebraic properties in relation to the inner product referred to in T1, e.g., $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$.

We have already seen an example of T2 in our analysis of the twin paradox (figure 5 on p. 26). Conceptually, T2 is similar to defining a line as the shortest path between two points, except that we define a geodesic as being the *longest* one (four our $+-$ signature).

2.4.3 Comparison of the systems

It is useful to compare the axiomatizations P, E, and T from sections 2.1.1-2.4.2 with each other in order to gain insight into how much “wobble room” there is in constructing theories of spacetime. Since they are logically equivalent, any statement occurring in one axiomatization can be proved as a theorem in any one of the others.

For example, we might wonder whether it is possible to equip Galilean spacetime with a metric. The answer is no, since a system with a metric would satisfy the axioms of system T, which are logically equivalent to our system P. The underlying reason for this is that in Galilean spacetime there is no natural way to compare the scales of distance and time.

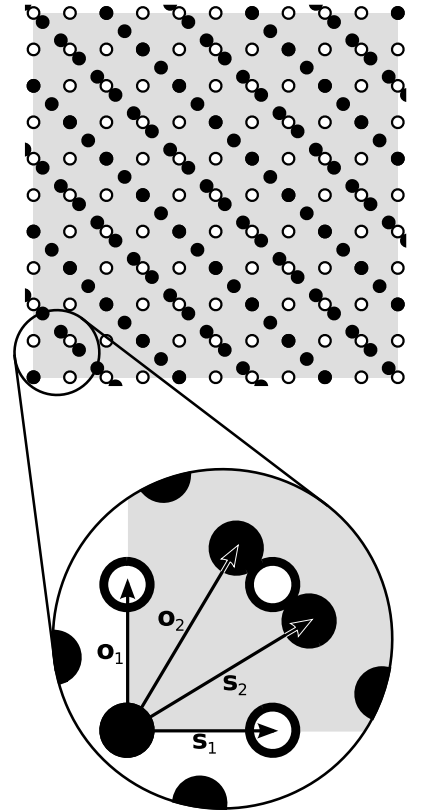
Or we could ask whether it is possible to compose variations on the theme of special relativity, alternative theories whose properties differ in some way. System P shows that this would be unlikely to succeed without violating the symmetry of spacetime.

Another interesting example is Amelino-Camelia’s doubly-special relativity,⁸ in which we have both an invariant speed c and an invariant length L , which is assumed to be the Planck length $\sqrt{\hbar G/c^3}$. The invariance of this length contradicts the existence of length contraction. In order to make his theory work, Amelino-Camelia is obliged to assume that energy-momentum vectors (section 4.3) have their own special inner product that violates the algebraic properties referred to in T2.

2.5 Lemma: spacetime area is invariant

In this section we prove from axioms P1-P6 that area in the $x-t$ plane is invariant, i.e., it does not change between frames of reference. This result was used in section 1.3.2 to find the form of the spacetime metric.

Consider figure f. Vectors \mathbf{o}_1 and \mathbf{s}_1 are orthogonal and have equal lengths as measured by a clock and a ruler (which are calibrated in units such that $c = 1$, e.g., seconds and light-seconds). The square lattice of white polka-dots is obtained from them by repeated addition. By assuming that this lattice construction is possible, we



f / Area is a scalar.

⁸arxiv.org/abs/gr-qc/0012051

are implicitly assuming postulate P2, flatness of spacetime.

The same properties hold for vectors \mathbf{o}_2 and \mathbf{s}_2 , which give the lattice of black dots. As required, the two lattices agree on their 45-degree diagonals. Now within the 10×10 portion of the white lattice shown with gray shading, we have an area of 100. In the same region we count about 100 or 101 black dots — there is some ambiguity because of the dots that lie on the boundary. The density of white and black dots is in fact exactly equal, as can be verified to any desired precision by making the region big enough. In other words, the diagram is drawn so that area is preserved, which is what we are going to show is required.

If it was observer 2 rather than 1 who was drawing the diagram, presumably she would choose to draw the black dots in a square lattice and vectors \mathbf{o}_2 and \mathbf{s}_2 at right angles. This would require vectors \mathbf{o}_1 and \mathbf{s}_1 to be opened up at an oblique angle and the white lattice to be non-square.

Now suppose we had *not* made area conserved. What if a region containing 100 white dots had held 200 black ones? Dot-counting is how the observers define area, so if this happened, they would have to agree that a boost by v , from frame 1 to frame 2, doubled the area of the gray region. Because spacetime is flat (P2) and homogeneous (P3), it is possible to take a geometrical shape inscribed in a certain region of spacetime and move, rotate, or flip it. And by isotropy of space (P3), a boost of velocity v is the same as a flip of the spatial dimension followed by a $-v$ boost and another flip. Area is conserved by a flip, so we find that a boost by $-v$, from frame 2 to frame 1, *also* doubles area. Thus a $+v$ boost followed by a $-v$ boost would cause a quadrupling of area. But a pair of equal and opposite boosts cancels out, so this is a contradiction. We conclude that if these symmetry principles hold, then spacetime area is the same for any two observers, so it is an invariant.

It may seem unnecessarily clumsy that we've used the idea of counting dots in the above argument, but remember that our main use of this result is to derive the form of the metric, and before the metric had been found, we had no system of measurement for relativity, so we had only very primitive techniques at our disposal.

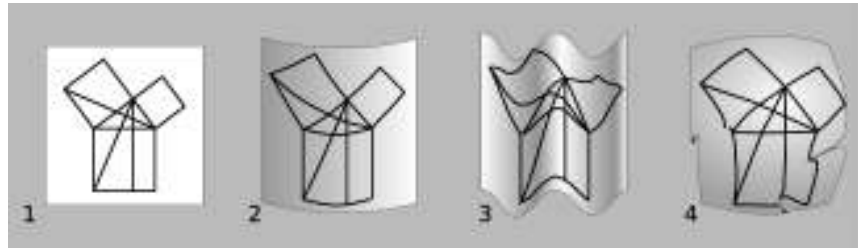
Problems

1 Section 2.5 gives an argument that spacetime area is a relativistic invariant. Is this argument also valid for Galilean relativity?

2 Section 2.5 gives an argument that spacetime area is a relativistic invariant. (a) Generalize this from 1+1 dimensions to 3+1. (b) Use this result to prove that there is no relativistic length contraction effect along an axis perpendicular to the velocity.

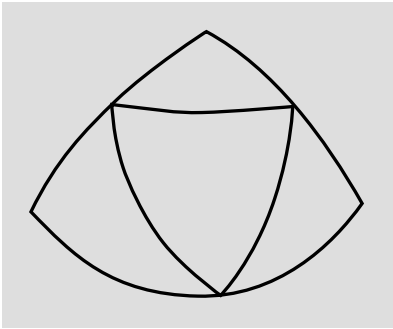
3 The purpose of this problem is to find how the direction of a physical object such as a stick changes under a Lorentz transformation. Part b of problem 2 shows that relativistic length contraction occurs only along the axis parallel to the motion. The generalization of the 1 + 1-dimensional Lorentz transformation to 2 + 1 dimensions therefore consists simply of augmenting equation (1) on p. 31 with $y' = y$. Suppose that a stick, in its own rest frame, has one end with a world-line $(\tau, 0, 0)$ and the other with (τ, p, q) , where τ is the stick's proper time. Call these ends A and B. In other words, we have a stick that goes from the origin to coordinates (p, q) in the (x, y) plane. Apply a Lorentz transformation for a boost with velocity v in the x direction, and find the equations of the world-lines of the ends of the stick in the new (t', x', y') coordinates. According to this new frame's notion of simultaneity, find the coordinates of B when A is at $(t', x', y') = (0, 0, 0)$. (a) In the special case where $q = 0$, recover the 1 + 1-dimensional result for length contraction given on p. 27. (b) Returning to the general case where $q \neq 0$, consider the angle θ that the stick makes with the x axis, and the related angle θ' that it makes with the x' axis in the new frame. Show that $\tan \theta' = \gamma \tan \theta$.

4 Section 2.2 discusses the idea that a two-dimensional bug living on the surface of a sphere could tell that its space was curved. Figure c on p. 45 shows one way of telling, by detecting the path-dependence of parallel transport. A different technique would be to look for violations of the Pythagorean theorem. In the figure below, 1 is a diagram illustrating the proof of the Pythagorean theorem in Euclid's *Elements* (proposition I.47). This diagram is equally valid if the page is rolled onto a cylinder, 2, or formed into a wavy corrugated shape, 3. These types of curvature, which can be achieved without tearing or crumpling the surface, are not real to the bug. They are simply side-effects of visualizing its two-dimensional universe as if it were embedded in a hypothetical third dimension — which doesn't exist in any sense that is empirically verifiable to the bug. Of the curved surfaces in the figure, only the sphere, 4, has curvature that the bug can measure; the diagram can't be plastered onto the sphere without folding or cutting and pasting. If a two-dimensional being lived on the surface of a cone, would it say that its space was curved, or not? What about a saddle shape?



Problem 4.

5 The discrepancy in parallel transport shown in figure c on p. 45 can also be interpreted as a measure of the triangle's angular defect d , meaning the amount $S - \pi$ by which the sum of its interior angles S exceeds the Euclidean value. (a) The figure suggests a simple way of verifying that the angular defect of a triangle inscribed on a sphere depends on area. It shows a large equilateral triangle that has been dissected into four smaller triangles, each of which is also approximately equilateral. Prove that $D = 4d$, where D is the angular defect of the large triangle and d the value for one of the four smaller ones. (b) Given that the proportionality to area $d = kA$ holds in general, find some triangle on a sphere of radius R whose area and angular defect are easy to calculate, and use it to fix the constant of proportionality k .



Remark: A being who lived on a sphere could measure d and A for some triangle and infer R , which is a measure of curvature. The proportionality of the effect to the area of the triangle also implies that the effects of curvature become negligible on sufficiently small scales. The analogy in relativity is that special relativity is a valid approximation to general relativity in regions of space that are small enough so that spacetime curvature becomes negligible.

Problem 5.

Chapter 3

Kinematics

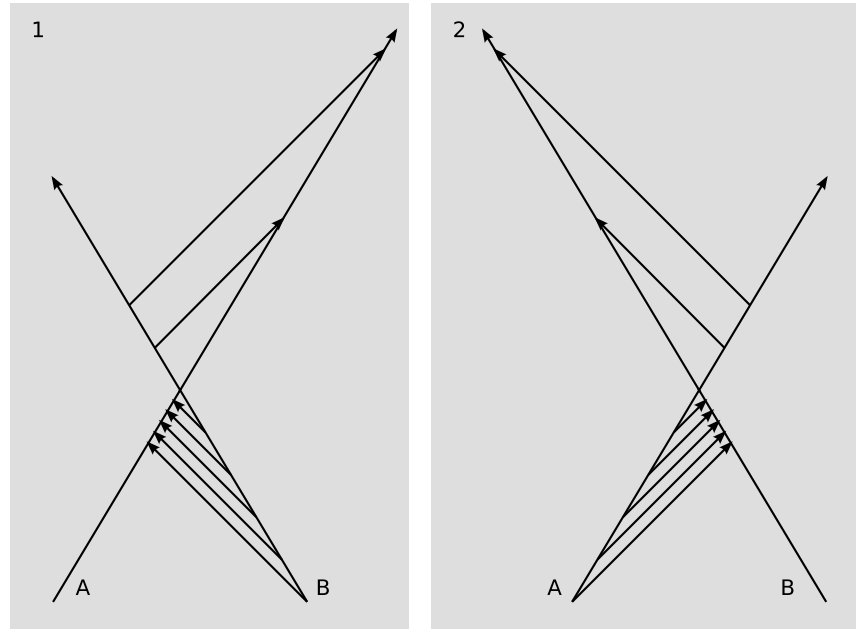
At this stage, many students raise the following questions, which turn out to be related to one another:

1. According to Einstein, if observers A and B aren't at rest relative to each other, then A says B's time is slow, but B says A is the slow one. How can this be? If A says B is slow, shouldn't B say A is fast? After all, if I took a pill that sped up my brain, everyone else would seem slow to me, and I would seem fast to them.
2. Suppose I keep accelerating my spaceship steadily. What happens when I get to the speed of light?
3. In all the diagrams in section 1.4, the parallelograms have their diagonals stretched and squished by a certain factor, which depends on v . What is the interpretation of this factor?

3.1 How can they both ... ?

Figure a shows how relativity resolves the first question. If A and B had an instantaneous method of communication such as Star Trek's subspace radio, then they could indeed resolve the question of who was really slow.

a / Signals don't resolve the dispute over who is really slow.

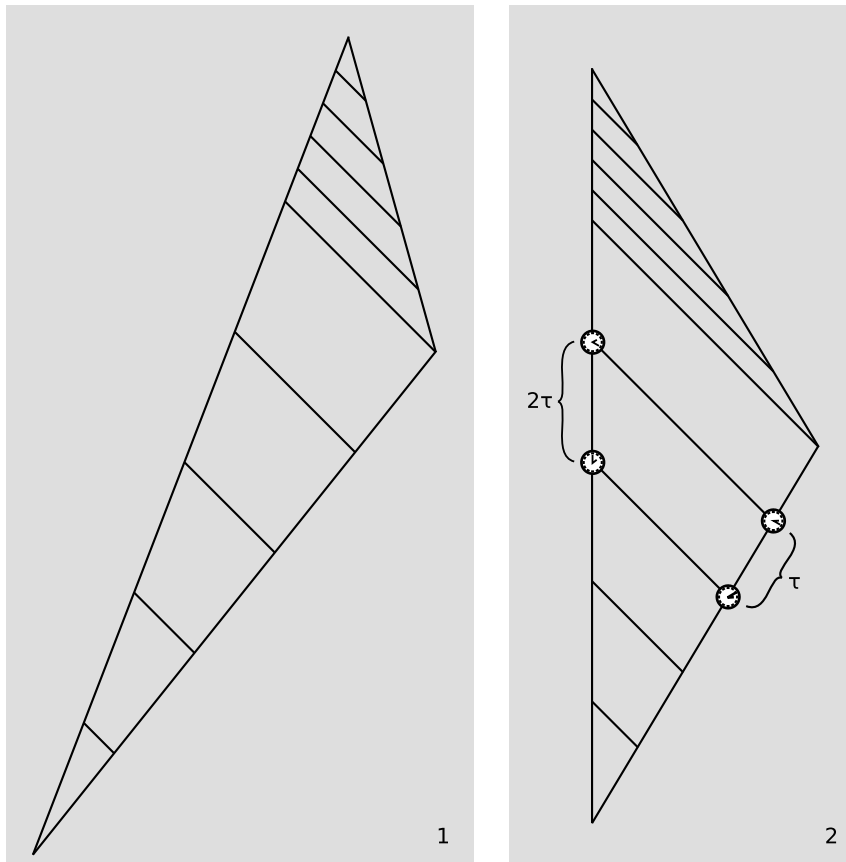


But relativity does not allow cause and effect to be propagated outside the light cone, so the best they can actually do is to send each other signals at c . In a/1, B sends signals to A at time intervals of one hour as measured by B's clock. According to A's clock, the signals arrive at an interval that is shorter than one hour as the two spaceships approach one another, then longer than an hour after they pass each other and begin to recede. As shown in a/2, the situation is entirely symmetric if A sends signals to B.

Who is really slow? Neither. If A, like many astronauts, cut her teeth as a jet pilot, it may occur to her to interpret the observations by analogy with the Doppler effect for sound waves. Figure a is in fact a valid diagram if the signals are clicks of sound, provided that we interpret it as being drawn in the frame of reference of the air. Sound waves travel at a fixed speed relative to the air, and the space and time units could be chosen such that the speed of *sound* was represented by a slope of ± 1 . But A will find that in the relativistic case, with signals traveling at c , her observations of the time intervals are not in quantitative agreement with the predictions she gets by plugging numbers into the familiar formulas for the Doppler shift of sound waves. She may then say, "Ah, the analogy with sound isn't quite right. I need to include a correction

factor for time dilation, since B's time is slow. I'm not slow, of course. I feel perfectly normal."

But her analogy is false and needlessly complicates the situation. In the version with sound waves and Galilean relativity, there are three frames of reference involved: A's, B's, and the air's. The relativistic version is simpler, because there are only two frames, A's and B's. It's neither helpful nor necessary to break down the observations into a factor describing what "really" happens and a correction factor to account for the relativistic distortions of "reality." All we need to worry about is the world-lines and intersections of world-lines shown in the spacetime diagrams, along with the metric, which allows us to compute how much proper time is experienced by each observer.



b / The twin paradox with signals sent back to earth by the traveling twin.

3.2 The stretch factor is the Doppler shift

Figure b shows how the ideas in the preceding section apply to the twin paradox. In b/1 we see the situation as described by an impartial observer, who says that both twins are traveling to the right. But even the impartial observer agrees that one twin's motion is inertial and the other's noninertial, which breaks the symmetry and

also allows the twins to meet up at the end and compare clocks. For convenience, b/2 shows the situation in the frame where the earthbound twin is at rest. Both panels of the figure are drawn such that the relative velocity of the twins is $3/5$, and in panel 2 this is the inverse slope of the traveling twin's world-lines. Straightforward algebra and geometry (problem 6, p. 76) shows that in this particular example, the period observed by the earthbound twin is increased by a factor of 2. But 2 is exactly the factor by which the diagonals of the parallelogram are stretched and compressed in a Lorentz transformation for a velocity of $3/5$. This is true in general: the stretching and squishing factors for the diagonals are the same as the Doppler shift. We notate this factor as D (which can stand for either "Doppler" or "diagonal"), and in general it is given by

$$D(v) = \sqrt{\frac{1+v}{1-v}}$$

(problem 7, p. 76).

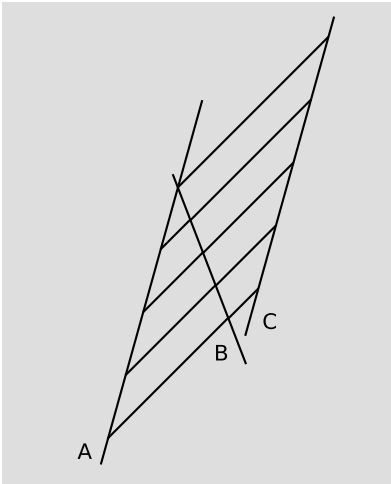
self-check A

If you measure with a ruler on figure b/2, you will find that the labeled sides of the quadrilateral differ by less than a factor of 2. Why is this?

▷ Answer, p. ??

This expression is for the longitudinal Doppler shift, i.e., the case where the source and observer are in motion directly away from one another (or toward one another if $v < 0$). In the purely transverse case, there is a Doppler shift $1/\gamma$ which can be interpreted as simply a measure of time dilation.

The useful identity $D(v)D(-v) = 1$ is trivial to prove algebraically, and has the following interpretation. Suppose, as in figure c, that A and C are at rest relative to one another, but B is moving relative to them. B's velocity relative to A is v , and C's relative to B is $-v$. At regular intervals, A sends lightspeed "pings" to B, who then immediately retransmits them to C. The interval between pings accumulates two Doppler shifts, and the result is their product $D(v)D(-v)$. But B didn't actually need to receive the original signal and retransmit it; the results would have been the same if B had just stayed out of the way. Therefore this product must equal 1, so $D(v)D(-v) = 1$.



c / Interpretation of the identity $D(v)D(-v) = 1$.

Ives-Stilwell experiments

Example 1

The transverse Doppler shift is a characteristic prediction of special relativity, with no nonrelativistic counterpart, and Einstein suggested it early on as a test of relativity. However, it is difficult to measure with high precision, because the results are sensitive to any error in the alignment of the 90-degree angle. Such experiments were eventually performed, with results that confirmed relativity,¹ but one-dimensional measurements provided both the

¹See, e.g., Hasselkamp, Mondry, and Scharmann, *Zeitschrift für Physik A: Hadrons and Nuclei* 289 (1979) 151.

earliest tests of the relativistic Doppler shift and the most precise ones to date. The first such test was done by Ives and Stilwell in 1938, using the following trick. The relativistic expression $D(v) = \sqrt{(1+v)/(1-v)}$ for the Doppler shift has the property that $D(v)D(-v) = 1$, which differs from the nonrelativistic result of $(1+v)(1-v) = 1 - v^2$. One can therefore accelerate an ion up to a relativistic speed, measure both the forward Doppler shifted frequency f_f and the backward one f_b , and compute $\sqrt{f_f f_b}$. According to relativity, this should exactly equal the frequency f_0 measured in the ion's rest frame.

In a particularly exquisite modern version of the Ives-Stilwell idea,² Saathoff et al. circulated Li^+ ions at $v = .064$ in a storage ring. An electron-cooler technique was used in order to reduce the variation in velocity among ions in the beam. Since the identity $D(v)D(-v) = 1$ is independent of v , it was not necessary to measure v to the same incredible precision as the frequencies; it was only necessary that it be stable and well-defined. The natural line width was 7 MHz, and other experimental effects broadened it further to 11 MHz. By curve-fitting the line, it was possible to achieve results good to a few tenths of a MHz. The resulting frequencies, in units of MHz, were:

$$\begin{aligned} f_f &= 582490203.44 \pm .09 \\ f_b &= 512671442.9 \pm 0.5 \\ \sqrt{f_f f_b} &= 546466918.6 \pm 0.3 \\ f_0 &= 546466918.8 \pm 0.4 \text{ (from previous experimental work)} \end{aligned}$$

The spectacular agreement with theory has made this experiment a lightning rod for anti-relativity kooks.

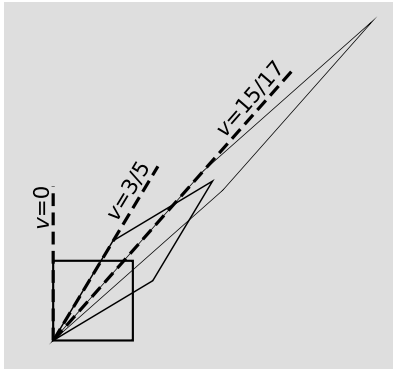
If one is searching for small deviations from the predictions of special relativity, a natural place to look is at high velocities. Ives-Stilwell experiments have been performed at velocities as high as 0.84, and they confirm special relativity.³

3.3 Combination of velocities

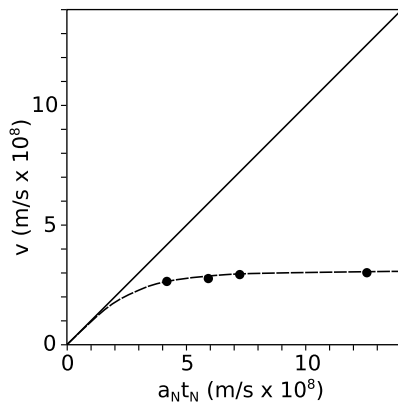
In nonrelativistic physics, velocities add in relative motion. For example, if a boat moves relative to a river, and the river moves relative to the land, then the boat's velocity relative to the land is found by vector addition. This linear behavior cannot hold relativistically. For example, if a spaceship is moving relative to the earth at velocity $3/5$ (in units with $c = 1$), and it launches a probe at velocity $3/5$ relative to itself, we can't have the probe moving at a velocity of $6/5$ relative to the earth, because this would be greater

²G. Saathoff et al., "Improved Test of Time Dilation in Relativity," Phys. Rev. Lett. 91 (2003) 190403. A publicly available description of the experiment is given in Saathoff's PhD thesis, www.mpi-hd.mpg.de/ato/homes/saathoff/diss-saathoff.pdf.

³MacArthur et al., Phys. Rev. Lett. 56 (1986) 282 (1986)



d / Two Lorentz transformations of $v = 3/5$ are applied one after the other. The transformations are represented according to the graphical conventions of section 1.4.



e / Example 2.

than the maximum speed of cause and effect, which is 1. To see how to add velocities relativistically, we consider the effect of carrying the two Lorentz transformations one after the other, figure d.

The inverse slope of the left side of each parallelogram indicates its velocity relative to the original frame, represented by the square. Since the left side of the final parallelogram has not swept past the diagonal, clearly it represents a velocity of *less* than 1, not more. To determine the result, we use the fact that the D factors multiply. We chose velocities $3/5$ because it gives $D = 2$, which is easy to work with. Doubling the long diagonal twice gives an over-all stretch factor of 4, and solving the equation $D(v) = 4$ for v gives the result, $v = 15/17$.

We can now see the answer to question 2 on p. 53. If we keep accelerating a spaceship steadily, we are simply continuing the process of acceleration shown in figure d. If we do this indefinitely, the velocity will approach $c = 1$ but never surpass it. (For more on this topic of going faster than light, see section 4.7.)

Accelerating electrons

Example 2

Figure e shows the results of a 1964 experiment by Bertozzi in which electrons were accelerated by the static electric field E of a Van de Graaff accelerator of length ℓ_1 . They were then allowed to fly down a beamline of length $\ell_2 = 8.4$ m without being acted on by any force. The time of flight t_2 was used to find the final velocity $v = \ell_2/t_2$ to which they had been accelerated. (To make the low-energy portion of the graph legible, Bertozzi's highest-energy data point is omitted.)

If we believed in Newton's laws, then the electrons would have an acceleration $a_N = Ee/m$, which would be constant if, as we pretend for the moment, the field E were constant. (The electric field inside a Van de Graaff accelerator is not really quite uniform, but this will turn out not to matter.) The Newtonian prediction for the time over which this acceleration occurs is $t_N = \sqrt{2m\ell_1/eE}$. An acceleration a_N acting for a time t_N should produce a final velocity $a_N t_N = \sqrt{2eV/m}$, where $V = E\ell_1$ is the voltage difference. (By conservation of energy, this equation holds even if the field is not constant.) The solid line in the graph shows the prediction of Newton's laws, which is that a constant force exerted steadily over time will produce a velocity that rises linearly and without limit.

The experimental data, shown as black dots, clearly tell a different story. The velocity asymptotically approaches a limit, which we identify as c . The dashed line shows the predictions of special relativity, which we are not yet ready to calculate because we haven't yet seen how kinetic energy depends on velocity at relativistic speeds. The calculation is carried out in example 4 on p. 85.

Note that the relationship between the first and second frames of reference in figure d is the same as the relationship between the second and third. Therefore if a passenger is to feel a steady sensation of acceleration (or, equivalently, if an accelerometer aboard the ship is to show a constant reading), then the *proper* time required to pass from the first frame to the second must be the same as the *proper* time to go from the second to the third. A nice way to express this is to define the *rapidity* $\eta = \ln D$. Combining velocities means multiplying D 's, which is the same as adding their logarithms. Therefore we can write the relativistic rule for combining velocities simply as

$$\eta_c = \eta_1 + \eta_2.$$

The passengers perceive the acceleration as steady if η increase by the same amount per unit of proper time. In other words, we can define a proper acceleration $d\eta/d\tau$, which corresponds to what an accelerometer measures.

Rapidity is convenient and useful, and is very frequently used in particle physics. But in terms of ordinary velocities, the rule for combining velocities can also be rewritten using identity [9] from section 3.6 as

$$v_c = \frac{v_1 + v_2}{1 + v_1 v_2}.$$

self-check B

How can we tell that this equation is written in natural units? Rewrite it in SI units.

▷ Answer, p. ??

3.4 No frame of reference moving at c

We have seen in section 3.3 that no continuous process of acceleration can boost a material object to c . That is, the subluminal (slower than light) nature of a electron or a person is a fundamental feature of its identity and can never be changed. Einstein can never get on his motorcycle and drive at c as he imagined when he was a young man, so we material beings can never see the world from a frame of reference that travels at c .

Our universe does, however, contain ingredients such as light rays, gluons, and gravitational waves that travel at c , so we might wonder whether these things could be put together to form observers who do move at c . But this is not possible according to special relativity, because if we let v approach infinity, extrapolation of figure d on p. 58 shows that the Lorentz transformation would compress all of spacetime onto the light cone, reducing its number of dimensions by 1. Distinct points would be merged, which would make it impossible to use this frame to describe the same phenomena that a subluminal observer could describe. That is, the transformation would not be one-to-one, and this is unacceptable physically.



f / A playing card returns to its original state when rotated by 180 degrees. Its orientation, unlike the orientation of an arrow, doesn't behave as a vector, since it doesn't transform in the usual way under rotations. Under a 180-degree rotation, a vector should negate itself rather than coming back to its original state.

3.5 The velocity and acceleration vectors

3.5.1 The velocity vector

In a freshman course in Newtonian mechanics, we would define a vector as something that has three components. Furthermore, we would require it to transform in a certain way under a rotation. For example, we could form the collection of numbers (e, T, DJIA) , where e is the fundamental charge, T is the temperature in Buffalo, New York, and DJIA measures how the stock market is doing. But this would not be a vector, since it doesn't act the right way when rotated (this particular “vector” is invariant under rotations). Figure f gives a less silly non-example. In contradistinction to a vector, a scalar is specified by a single real number and is invariant under rotations.

The most basic example of a Newtonian vector was a displacement $(\Delta x, \Delta y, \Delta z)$, and from the displacement vector we would go on to construct other quantities such as a velocity vector $\mathbf{v} = \Delta \mathbf{r} / \Delta t$. This worked because in Newtonian mechanics Δt was treated as a scalar, and dividing a vector by a scalar produces something that again transforms in the right way to be a vector.

Now let's upgrade to relativity, and work through the same steps by analogy. When I say “vector” in this book, I mean something that in 3+1 dimensions has four components. This can also be referred to as a four-vector. Our only example so far has been the spacetime displacement vector $\Delta \mathbf{r} = (\Delta t, \Delta x, \Delta y, \Delta z)$. This vector transforms according to the Lorentz transformation. In general, we require as part of the definition of a (four-)vector that it transform in the usual way under both rotations and boosts (Lorentz transformations). We might now imagine that the next step should be to construct a velocity four-vector $\Delta \mathbf{r} / \Delta t$. But relativistically, the quantity $\Delta \mathbf{r} / \Delta t$ would not transform like a vector, e.g., if \mathbf{r} was spacelike, then there would be a frame in which we had $\Delta t = 0$, and then $\Delta \mathbf{r} / \Delta t$ would be finite in some frames but infinite in others, which is absurd.

To construct a valid vector, we have to divide $\Delta \mathbf{r}$ by a scalar. The only scalar that could be relevant would be the *proper* time $\Delta \tau$, and this is indeed how the velocity vector is defined in relativity. For an inertial world-line (one with constant velocity), we define $\mathbf{v} = \Delta \mathbf{r} / \Delta \tau$. The generalization to noninertial world-lines requires that we make this definition into a derivative:

$$\mathbf{v} = \frac{d\mathbf{r}}{d\tau}$$

Not all objects have well-defined velocity vectors. For example, consider a ray of light with a straight world-line, so that the derivative $d.../d...$ is the same as the ratio of finite differences $\Delta.../\Delta...$, i.e., calculus isn't needed. A ray of light has $v = c$, so that applying the metric to any segment of its world-line gives

$\Delta\tau = 0$. Attempting to calculate $\mathbf{v} = \Delta\mathbf{r}/\Delta\tau$ then gives something with infinite components. We will see in section 4.3.4 that *all* massless particles, not just photons, travel at c , so the same would apply to them. Therefore a velocity vector is only defined for particles whose world-lines are timelike, i.e., massive particles.

Velocity vector of an object at rest

Example 3

An object at rest has $\mathbf{v} = (1, 0)$. The first component indicates that if we attach a clock to the object with duct tape, the proper time measured by the clock suffers no time dilation according to an observer in this frame, $dt/d\tau = 1$. The second component tells us that the object's position isn't changing, $dx/d\tau = 0$.

3.5.2 The acceleration vector

The acceleration vector is defined as the derivative of the velocity vector with respect to proper time,

$$\mathbf{a} = \frac{d\mathbf{v}}{d\tau}.$$

It measures the curvature of a world-line. Its squared magnitude is the minus the square of the *proper acceleration*, meaning the acceleration that would be measured by an accelerometer carried along that world-line. The proper acceleration is only approximately equal to the magnitude of the Newtonian acceleration three-vector, in the limit of small velocities.

Constant proper acceleration

Example 4

▷ Suppose a spaceship moves so that the acceleration is judged to be the constant value a by an observer on board. Find the motion $x(t)$ as measured by an observer in an inertial frame.

▷ Let τ stand for the ship's proper time, and let dots indicate derivatives with respect to τ . The ship's velocity has magnitude 1, so

$$\dot{t}^2 - \dot{x}^2 = 1.$$

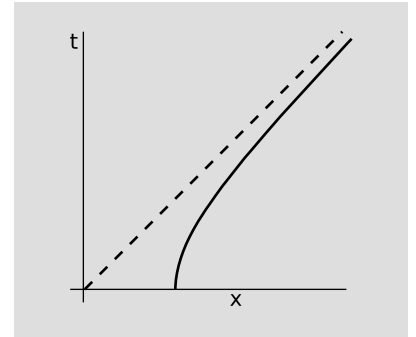
An observer who is instantaneously at rest with respect to the ship judges is to have an acceleration vector $(0, a)$ (because the low-velocity limit applies). The observer in the (t, x) frame agrees on the magnitude of this vector, so

$$\ddot{t}^2 - \ddot{x}^2 = -a^2.$$

The solution of these differential equations is $t = \frac{1}{a} \sinh a\tau$, $x = \frac{1}{a} \cosh a\tau$ (choosing constants of integration so that the expressions take on their simplest forms). Eliminating τ gives

$$x = \frac{1}{a} \sqrt{1 + a^2 t^2},$$

shown in figure g. The world-line is a hyperbola, and this type of motion is sometimes referred to as hyperbolic motion.



g / A spaceship (curved world-line) moves with an acceleration perceived as constant by its passengers.

As t approaches infinity, dx/dt approaches the speed of light. In the same limit, x increases exponentially with proper time, so that surprisingly large distances can in theory be traveled within a human lifetime (problem 7, p. 112). Some further properties of hyperbolic motion are developed in problems 10, 11, and 12.

Another interesting feature of this problem is the dashed-line asymptote, which is lightlike. Suppose we interpret this as the world-line of a ray of light. The ray comes closer and closer to the ship, but will never quite catch up. Thus provided that the rocket never stops accelerating, the entire region of spacetime to the left of the dashed line is forever hidden from its passengers. That is, an observer who undergoes constant acceleration has an *event horizon* — a boundary that prevents her from observing anything on the other side. You may have heard about the event horizon associated with a black hole. This example shows that we can have event horizons even when there is no gravity at all.

3.5.3 Constraints on the velocity and acceleration vectors

Counting degrees of freedom

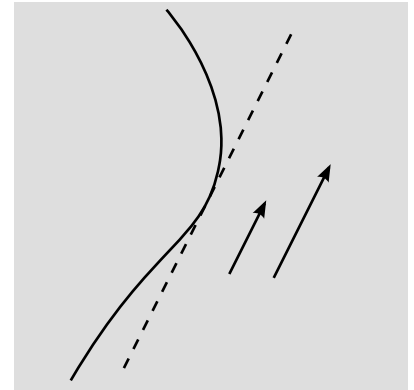
There is something misleading about the foregoing treatment of the velocity and acceleration vectors, and the easiest way to see this is by introducing the idea of a *degree of freedom*. Often we can describe a system using a list of real numbers. For the hand on a clock, we only need one number, such as 3 o'clock. This is because the hand is constrained to stay in the plane of the clock's face and also to keep its tail at the center of the circle. Since one number describes its position, we say that it has one degree of freedom. If a hiker wants to know where she is on a map, she has two degrees of freedom, which could be specified as her latitude and longitude. If she was in a helicopter, there would be no constraint to stay on the earth's surface, and the number of degrees of freedom would be increased to three. If we also considered the helicopter's velocity to be part of the description of its state, then there would be a total of six degrees of freedom: one for each coordinate and one for each component of the velocity vector.

Now suppose that we want to specify a particle's velocity and acceleration. In Newtonian mechanics, we would describe these three-vectors as possessing a total of six degrees of freedom: v_x , v_y , v_z , a_x , a_y , and a_z . Upgrading from Newtonian mechanics to relativity can't change the number of degrees of freedom. For example, an electron's acceleration is fully determined by the force we exert on it, and we might control that acceleration by placing a proton nearby and producing an electrical attraction. The position of the proton (three degrees of freedom for its three coordinates) determines the electron's acceleration, so the acceleration has exactly three degrees of freedom as well.

This means that there must be some hidden redundancy in the *eight* components of the velocity and acceleration four-vectors. The system only has six degrees of freedom, so there must be *two constraints* that we didn't know about. Similarly, I've gone hiking and had my GPS unit claim that I was a thousand feet above a lake or three thousand feet under a mountain. In those situations there was a constraint that I knew about but that the GPS didn't: that I was on the surface of the earth.

Normalization of the velocity

The first constraint arises naturally from a geometrical interpretation of the velocity four-vector, shown in figure h. The curve represents the world-line of a particle. The dashed line is drawn tangent to the world-line at a certain moment. Under a microscope, the dashed line, which represents a possible inertial motion of a particle, is indistinguishable from the solid curve, which is noninertial. The dashed line has a slope $\Delta t/\Delta x = 2$, which corresponds to a velocity $\Delta x/\Delta t = 1/2$. The figure is drawn in $1 + 1$ dimensions, but in $3 + 1$ dimensions we would want to know more than this number. We would want to know the orientation of the dashed line in the three spatial dimensions, i.e., not just the speed of the particle but also its direction of motion. All the desired information can be encapsulated in a vector. Both of the vectors shown in the figure are parallel to the dashed line, so even though they have different lengths, there is no difference between the velocities they represent. Since we want the particle to have a single well-defined vector to represent its velocity, we want to pick one vector from among all the vectors parallel to the dashed line, and call that “the” velocity vector.



h / Both vectors are tangent vectors.

We have already implicitly made this choice. It follows from the original definition $\mathbf{v} = d\mathbf{r}/d\tau$ that the velocity vector's squared magnitude $v^2 = \mathbf{v} \cdot \mathbf{v}$ is always equal to 1, even though the object whose motion it describes is not moving at the speed of light. This, along with the requirement that the velocity vector lie within the future rather than the past light cone, uniquely specifies which tangent vector we want. The requirement $v^2 = 1$ is an example of a recurring idea in physics and mathematics called normalization. The idea is that we have some object (a vector, a function, ...) that could be scaled up or down by any amount, but from among all the possible scales, there is only one that is the right one. For example, a gambler might place a horse's chance of winning at 9 to 1, but a physicist would divide these by 10 in order to normalize the probabilities to 0.9 and 0.1, the idea being that the total probability should add up to 1. Our definition of the velocity vector implies that it is normalized. Thus an alternative, geometrical definition of the velocity vector would have been that it is the vector that is tangent to the particle's world-line, future-directed, and normalized to 1.

When we hear something referred to as a “vector,” we usually take this is a statement that it not only transforms as a vector, but also that it adds as a vector. But the sum of two velocity vectors would not typically be a valid velocity vector at all, since it would not have unit magnitude. This lack of additivity would in any case have been expected because velocities don’t add linearly in relativity (section 3.3).

self-check C

Velocity vectors are required to have $v^2 = 1$. If a vector qualifies as a valid velocity vector in some frame, could it be invalid in another frame?

▷ Answer, p. ??

A nice way of thinking about velocity vectors is that every such vector represents a potential observer. That is, the velocity vectors are the observer-vectors \mathbf{o} of chapter 1, but with a normalization requirement $o^2 = 1$ that we did not impose earlier. An observer writes her own velocity vector as $(1, 0)$, i.e., as the unit vector in the timelike direction. Since we have no notion of adding one observer to another observer, it makes sense that velocity vectors don’t add relativistically. Similarly, there is no meaningful way to define the magnitude of an observer, so it makes sense that the magnitude of a velocity vector carries no useful information and can arbitrarily be set equal to 1.

Regarding the magnitude, note also that the magnitude of a vector is frame-invariant, and therefore it wouldn’t make sense to imagine that the magnitude of an object’s four-velocity would produce some number telling you how fast the object was going. How fast relative to what?

If \mathbf{u} and \mathbf{v} are both future-directed, properly normalized velocity vectors, and if the signature is $+- --$ as in this book, then their inner product is $\gamma = \mathbf{u} \cdot \mathbf{v}$, the gamma factor, introduced in section 1.3.3, p. 25, corresponding their relative velocity.

Orthogonality of the velocity and acceleration

Now for the second of the two constraints deduced on p. 62.

Suppose an observer claims that at a certain moment in time, a particle has $\mathbf{v} = (1, 0)$ and $\mathbf{a} = (3, 0)$. That is, the particle is at rest ($v_x = 0$) and its v_t is growing by 3 units per second. This is impossible, because after an infinitesimal time interval dt , this rate of change will result in $\mathbf{v} = (1 + 3 dt, 0)$, which is not properly normalized: its magnitude has grown from 1 to $1 + 3 dt$. The observer is mistaken. This is not a possible combination of velocity and acceleration vectors. In general (problem 9, p. 76), we always have the following constraint on the velocity and acceleration vectors:

$$\mathbf{a} \cdot \mathbf{v} = 0.$$

This is analogous to the three-dimensional idea that in uniform cir-

cular motion, the perpendicularity of the velocity and acceleration three-vectors is what causes the velocity vector to rotate without changing its magnitude.

3.6 ★ Some kinematic identities

In addition to the relations

$$D(v) = \sqrt{\frac{1+v}{1-v}} \quad \text{and} \\ v_c = \frac{v_1 + v_2}{1 + v_1 v_2},$$

the following identities can be handy. If stranded on a desert island you should be able to rederive them from scratch. Don't memorize them.

$$\begin{array}{lll} [1] & v = (D^2 - 1)/(D^2 + 1) & [5] \quad \eta = \ln D \quad [10] \quad D_c = D_1 D_2 \\ [2] & \gamma = (D^{-1} + D)/2 & [6] \quad v = \tanh \eta \quad [11] \quad \eta_c = \eta_1 + \eta_2 \\ [3] & v\gamma = (D - D^{-1})/2 & [7] \quad \gamma = \cosh \eta \quad [12] \quad v_c \gamma_c = (v_1 + v_2)\gamma_1 \gamma_2 \\ [4] & D(v)D(-v) = 1 & [8] \quad v\gamma = \sinh \eta \\ & & [9] \quad \tanh(x + y) = \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y} \end{array}$$

The hyperbolic trig functions are defined as follows:

$$\begin{aligned} \sinh x &= \frac{e^x - e^{-x}}{2} \\ \cosh x &= \frac{e^x + e^{-x}}{2} \\ \tanh x &= \frac{\sinh x}{\cosh x} \end{aligned}$$

Their inverses are built in to some calculators and computer software, but they can also be calculated using the following relations:

$$\begin{aligned} \sinh^{-1} x &= \ln \left(x + \sqrt{x^2 + 1} \right) \\ \cosh^{-1} x &= \ln \left(x + \sqrt{x^2 - 1} \right) \\ \tanh^{-1} x &= \frac{1}{2} \ln \left(\frac{1+x}{1-x} \right) \end{aligned}$$

Their derivatives are, respectively, $(x^2 + 1)^{-1/2}$, $(x^2 - 1)^{-1/2}$, and $(1 - x^2)^{-1}$.

3.7 ★ The projection operator

A frequent source of confusion in relativity is that we write down equations that are coordinate-dependent, but forget the dependency. Similarly, it is possible to write expressions that are only valid for one choice of signature. The following notation, defining a projection operator P , is one tool for avoiding these difficulties.

$$P_{\mathbf{o}}\mathbf{r} = \mathbf{r} - \frac{\mathbf{r} \cdot \mathbf{o}}{\mathbf{o} \cdot \mathbf{o}}\mathbf{o} \quad (1)$$

Usually \mathbf{o} is the future timelike vector representing a certain observer, but the definition can be applied as long as \mathbf{o} isn't lightlike. The idea being expressed is that we want to get rid of any part of \mathbf{r} that is parallel to \mathbf{o} 's arrow of time. In a graph constructed according to \mathbf{o} 's Minkowski coordinates, we cast \mathbf{r} 's shadow down perpendicularly onto the spacelike axis, or the spacelike three-plane in $3+1$ dimensions. This is why P is referred to as a projection operator. The notation sometimes allows us to express the things that we would otherwise express by explicitly or implicitly constructing and referring to \mathbf{o} 's spacelike Minkowski coordinates. P has the following properties:

1. $\mathbf{o} \cdot P_{\mathbf{o}}\mathbf{r} = 0$
2. $\mathbf{r} - P_{\mathbf{o}}\mathbf{r}$ is parallel to \mathbf{o} .
3. $P_{\mathbf{o}}\mathbf{o} = 0$
4. $P_{\mathbf{o}}P_{\mathbf{o}}\mathbf{r} = P_{\mathbf{o}}\mathbf{r}$
5. $P_{c\mathbf{o}} = P_{\mathbf{o}}$
6. $P_{\mathbf{o}}$ is linear, i.e., $P_{\mathbf{o}}(\mathbf{q} + \mathbf{r}) = P_{\mathbf{o}}\mathbf{q} + P_{\mathbf{o}}\mathbf{r}$ and $P_{\mathbf{o}}(c\mathbf{r}) = cP_{\mathbf{o}}\mathbf{r}$
7. $\frac{d}{dx}P_{\mathbf{o}}\mathbf{r} = P_{\mathbf{o}}\frac{d\mathbf{r}}{dx}$, where x is any variable and \mathbf{o} doesn't depend on x .
8. If \mathbf{o} and \mathbf{v} are both future timelike, and $|\mathbf{o}|^2 = 1$, then we can express \mathbf{v} as $\mathbf{v} = P_{\mathbf{o}}\mathbf{v} + \gamma\mathbf{o}$, where γ has the usual interpretation for world-lines that coincide with these two vectors.

All of these hold regardless of whether the signature is $+- - -$ or $- + + +$, and none of them refer to any coordinates. Properties 1 and 2 can serve as an alternative, geometrical definition of P . Property 3 says that an observer considers herself to be at rest. 4 is a general property of all projection operators. 8 splits the vector into its spatial and temporal parts according to \mathbf{o} .

Sometimes if we know a position, velocity, or acceleration four-vector, we want to find out how these would be measured by a particular observer using clocks and rulers. The following table shows

how to switch back and forth between the two representations. We use, for example, the notation \mathbf{v}_o to mean the velocity vector of the form $(0, v_x, v_y, v_z)$ that would be measured by an observer whose velocity vector is \mathbf{o} (so that the subscript is an “o” for “observer,” not a zero). Since this type of vector, expressed in the Minkowski coordinates of observer \mathbf{o} , has a zero time component, we refer to it as a three-vector. In all of these expressions, the velocity vectors \mathbf{o} and \mathbf{v} are assumed to be normalized, and the signature is assumed to be $+- - -$ (one implication being that $\mathbf{o} \cdot \mathbf{v}$ is simply γ).

<i>finding the three-vector from the four-vector</i>	<i>finding the four-vector from the three-vector</i>
$\mathbf{x}_o = P_o \mathbf{x}$ $\mathbf{v}_o = \frac{P_o \mathbf{v}}{\mathbf{o} \cdot \mathbf{v}}$ $\mathbf{a}_o = \frac{1}{(\mathbf{o} \cdot \mathbf{v})^2} [P_o \mathbf{a} - (\mathbf{o} \cdot \mathbf{a}) \mathbf{v}_o]$	$\mathbf{v} = \gamma(\mathbf{o} + \mathbf{v}_o)$ $\mathbf{a} = \gamma^3(\mathbf{a}_o \cdot \mathbf{v}_o) \mathbf{v} + \gamma^2 \mathbf{a}_o$, where \mathbf{v} is found as above

As an example of how these are derived, the three-velocity \mathbf{v}_o is the derivative of \mathbf{x}_o with respect to observer \mathbf{o} 's Minkowski time coordinate t , whereas the four-velocity is defined as the derivative of \mathbf{x} with respect to the proper time τ of the world-line being observed. Therefore we have

$$\begin{aligned} \mathbf{v}_o &= \frac{d\mathbf{x}_o}{dt} \\ &= \frac{dP_o \mathbf{x}}{dt} \end{aligned}$$

and applying property 7 of the projection operator this becomes

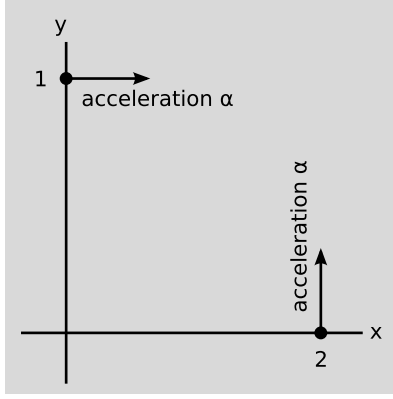
$$\begin{aligned} \mathbf{v}_o &= P_o \frac{d\mathbf{x}}{dt} \\ &= P_o \frac{d\mathbf{x}}{d\tau} \frac{d\tau}{dt} \\ &= \frac{1}{\gamma} P_o \frac{d\mathbf{x}}{d\tau} \\ &= \frac{1}{\mathbf{o} \cdot \mathbf{v}} P_o \frac{d\mathbf{x}}{d\tau} \\ &= \frac{P_o \mathbf{v}}{\mathbf{o} \cdot \mathbf{v}}. \end{aligned}$$

The similar but messier derivation of the expression for \mathbf{a}_o is problem 15. In manipulating expressions of this type, the identity $d\gamma/dt = \gamma^3 \mathbf{a}_o \cdot \mathbf{v}_o$ is often handy (problem 14).

Lewis-Tolman paradox

Example 5

The following example is a form of a paradox discussed by Lewis and Tolman in 1909. Figure i shows the frame of reference of observer \mathbf{o} in which identical particles 1 and 2 are at initially rest and located at equal distances ℓ from the origin along the y and x axes. External forces of equal strength act in the directions shown



i / Example 5.

by the arrows so as to produce accelerations of magnitude α . The system is in rotational equilibrium $dL/dt = 0$, because the rate at which particle 1 picks up clockwise angular momentum is the same as the rate at which 2 acquires it in the counterclockwise direction.

Now change to the frame of reference \mathbf{o}' , moving to the right relative to \mathbf{o} at velocity v . Particle 2's distance from the origin is Lorentz-contracted from ℓ to ℓ/γ , so its angular momentum is also reduced by $1/\gamma$. It now appears that the system's total angular momentum is increasing in the clockwise sense. How can we have rotational equilibrium in one frame, but not another?

The resolution of the paradox is that the accelerations transform as well. In the original frame \mathbf{o} , the four-velocities are $\mathbf{v}_1 = \mathbf{v}_2 = (1, 0, 0, 0)$, and the four-accelerations are $\mathbf{a}_1 = (0, \alpha, 0, 0)$ and $\mathbf{a}_2 = (0, 0, \alpha, 0)$. Applying a Lorentz transformation, we have $\mathbf{v}'_1 = \mathbf{v}'_2 = (\gamma, -\gamma v, 0, 0)$ and

$$\begin{aligned}\mathbf{a}'_1 &= \alpha(-\gamma v, \gamma, 0, 0) \\ \mathbf{a}'_2 &= \alpha(0, 0, 1, 0).\end{aligned}$$

Our definition of angular momentum is expressed in terms of *three*-vectors such as $\mathbf{a}_{\mathbf{o}'1}$ and $\mathbf{a}_{\mathbf{o}'2}$, not four-vectors like \mathbf{a}'_1 and \mathbf{a}'_2 . We have

$$\frac{dL'}{dt'} = m\mathbf{a}_{\mathbf{o}'1x}\ell - m\mathbf{a}_{\mathbf{o}'2y}\frac{\ell}{\gamma}.$$

Using the relations $\mathbf{v}_{\mathbf{o}} = \gamma^{-1}P_{\mathbf{o}}\mathbf{v}$ and $\mathbf{a}_{\mathbf{o}} = \gamma^{-2}[P_{\mathbf{o}}\mathbf{a} - (\mathbf{o} \cdot \mathbf{a})\mathbf{v}_{\mathbf{o}}]$, we find

$$\begin{aligned}v_{\mathbf{o}'1x} &= -v, \\ a_{\mathbf{o}'1x} &= \frac{1}{\gamma^2}[\alpha\gamma - (-\alpha\gamma v)(-v)] = \frac{\alpha}{\gamma^3},\end{aligned}$$

and

$$a_{\mathbf{o}'2y} = \frac{\alpha}{\gamma^2}.$$

The result is

$$\frac{dL'}{dt'} = m\frac{\alpha}{\gamma^3}\ell - m\frac{\alpha}{\gamma^2}\frac{\ell}{\gamma},$$

which is zero.

3.8 ★ Faster-than-light frames of reference?

We recall from section 3.4 that special relativity doesn't permit the existence of observers who move at c . This is because if two observers differ in velocity by c , then the Lorentz transformation between them is not a one-to-one map, which is physically unacceptable.

But what about a *superluminal* observer, one who moves faster than c ? With charming naivete, the special-effects technicians for Star Trek attempted to show the frame of reference of such an observer in scenes where a field of stars rushed past the Enterprise. (Never mind that the stars, which pass in front of and behind the spaceship, should actually be a million times larger than it.) Actually such an observer would consider her own world-line, which we call spacelike, to be timelike, while the world-line of a star such as our sun, which we consider timelike, would be spacelike in her opinion. Our sun's world-line might, for example, be orthogonal to hers, in which case the sun would not appear to her as an object in motion but rather as a line stretching across space, which would wink into existence and then wink back out. A typical transformation between our frame and the frame of such an observer would be the map S defined by $(t', x') = (x, t)$, simply swapping the time and space coordinates. The "swap" transformation S is one-to-one, and therefore not subject to the objection raised previously to frames moving at c . S happens to be a boost by an infinite velocity, but we can also obtain boosts for velocities $c < v < \infty$ and $-\infty < v < -c$ by combining S with a (subluminal) Lorentz transformation; given a superluminal world-line ℓ , we first transform into a frame in which ℓ is a line of simultaneity, and then we apply S .

But this was all in 1+1 dimensions. In 3+1 dimensions, what is the equivalent of S ? One possibility is something like $(t', x', y', z') = (x, t, t, t)$, but this isn't one-to-one. We can't squish three dimensions to one or expand one to three without merging points or splitting one point into many.

Another possibility would be a one-to-one transformation such as $(t', x', y', z') = (x, t, y, z)$. The trouble with this version is that it violates the isotropy of spacetime (section 2.3, p. 46). For example, consider the vector $(1, 0, 1, 0)$ in the unprimed coordinates. This lies on the light cone, and could point along the world-line of a ray of light. After the transformation to the primed coordinates, this vector becomes $(0, 1, 1, 0)$, which points along a line of simultaneity. The primed observer says that the speed of light in this direction is infinite, and yet there are other directions in which it has a finite value. This clearly violates isotropy.

A surprisingly large number of papers, going all the way back to the birth of relativity, have been written by people trying to find a way to extend the Lorentz transformations to superluminal speeds,

and these have all turned out to be failures. In fact, there are no-go theorems showing that there can be no such thing as a superluminal observer in our 3+1-dimensional universe.^{4,5}

The nonexistence of FTL *frames* does not immediately rule out the possibility of FTL *motion*. (After all, we do have motion at c , but no frames moving at c .) For more about faster-than-light motion in relativity, see section 4.7, p. 107.

3.9 ★ Thickening of a curve

3.9.1 A geometrical interpretation of the acceleration

We've interpreted the acceleration vector as a measure of the curvature of a world-line, but to make this more than a tool for visualization, we would have to define what we mean by curvature. A good way to approach this is shown in figure j/1. Here a circle of circumference L has been expanded, like a loaf of rising bread, to a circle of greater circumference L^* . This increase is only because the circle is curved. If we do the same thing with a line segment, j/2, there is no increase in length. The increase in the length tells us about the curvature.

Quantitatively, suppose that the thickness of the shaded area is Δh . Then the increase in circumference $\Delta L = L^* - L$ is given by

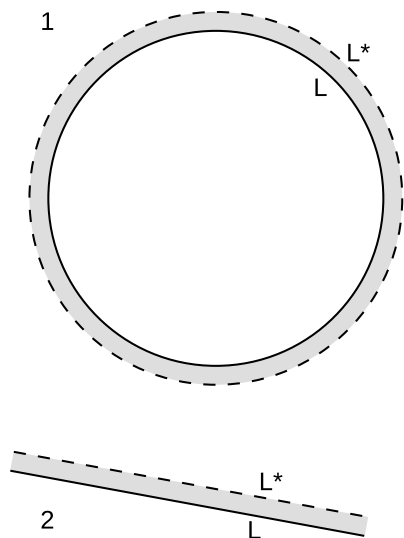
$$\frac{1}{L} \frac{\Delta L}{\Delta h} = k,$$

where k is a measure of curvature, and $k = 1/r$ for a circle. We can take this as a definition of the curvature of a curve embedded in a two-dimensional Euclidean plane. The curves in figure j/1 both have constant curvature, and if we had applied our definition to any short segment of them, we would have gotten the same answer. For a curve with varying curvature, such as a letter "S," the curvature can be defined as the appropriate limit at any given point, as the length of the segment enclosing the point approaches zero. Note that we had to pick an orientation for the expansion, i.e., a direction in which to expand. Given this orientation, it makes sense to talk about signed values of h and k . If we choose the outward orientation for a circle, then its k is positive.

An interesting point about this definition is that it is *extrinsic* rather than *intrinsic*, in the sense defined in section 2.2.1, p. 45. That is, it depends on how the curve is embedded in the ambient two-dimensional space, and it depends on the Euclidean metric of that space. Because a curve is a one-dimensional object, there

⁴Gorini, "Linear Kinematical Groups," Commun. Math. Phys. 21 (1971) 150. Open access via Project Euclid at projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.cmp/1103857292.

⁵Andreka et al., "A logic road from special relativity to general relativity," arxiv.org/abs/1005.0960, theorem 2.1.



j / One-sided thickenings of a circle and a line segment.

is nothing internal to the curve that would allow us to define its curvature. Imagine yourself as a tiny bug — so tiny that you are pointlike. If the curve represents your universe, then you can explore it as much as you like, but you can never detect any internal evidence of its curvature. This is not the case in two dimensions. For example, a bug living on the two-dimensional surface of a sphere can detect its curvature by drawing triangles and measuring how much the sum of their interior angles differs from 180 degrees. This would be an *intrinsic* measure of curvature. (See problem 4, p. 51.)

The definition given above is readily extended from Euclidean space to $1 + 1$ dimensions of spacetime. Figure k shows a one-sided thickening of an accelerated world-line. Although the shaded area doesn't look uniformly thick to our Euclidean eyes, it is. For example, each of the dotted lines is orthogonal to the original world-line on the left, and they all have the same length Δh as measured by an observer who traces that world-line. That is, each of these lines could represent a rigid measuring rod carried by that observer, drawn along a line that that observer considers to be a line of simultaneity at that time. In analogy to the Euclidean case, we have

$$\frac{1}{\tau} \frac{\Delta \tau}{\Delta h} = \frac{1}{a}.$$

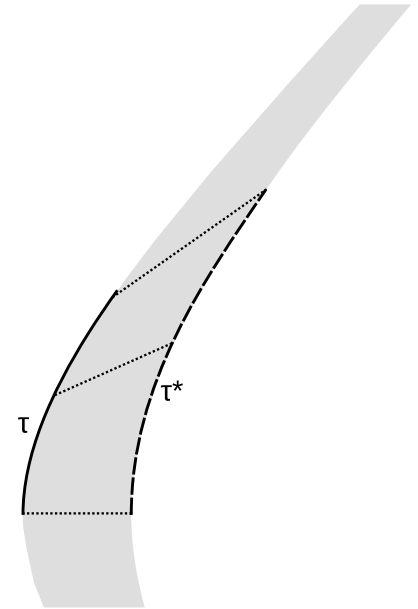
3.9.2 Bell's spaceship paradox

A variation on the situation shown in figure k leads to a paradox with philosophical implications proposed by John Bell. Bell went around the CERN cafeteria proposing the following thought experiment to the physicists eating lunch, and he found that nearly all of them got it wrong. Let two spaceships accelerate as shown in figure l. Each ship is equipped with a yard-arm, and a thread is tied between the two arms, l/1. Unaccelerated observer \mathbf{o} uses Minkowski coordinates (t, x) , as shown in l/2. The accelerations, as judged by \mathbf{o} , are equal for the two ships as functions of t . Does the thread break, due to Lorentz contraction?

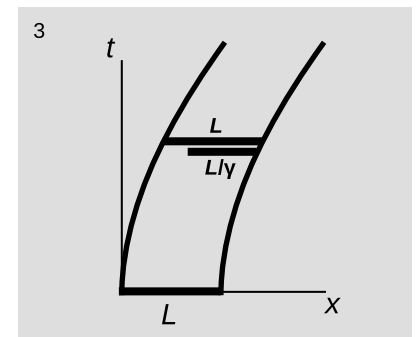
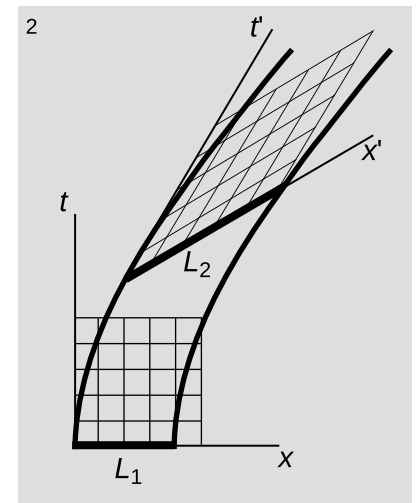
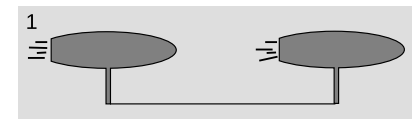
A crucial difference between figures k and l/2 is that in the former, the thickening of the world-line has been carried out along the dotted normals, whereas the latter the second world-line is simply a copy of the first that has been shifted to the right, parallel to the x axis.

The popular answer in the CERN cafeteria was that the thread would not break, the reasoning being that Lorentz contraction is a frame-dependent effect, and no such contraction would be observed in the rockets' frames.

The error in this reasoning is that the accelerations of the two ships were specified to be equal in frame \mathbf{o} , not in the frames of the rockets. The Minkowski coordinates (t', x') shown in l/2 correspond to the frame of an inertial observer \mathbf{o}' who is momentarily moving



k / One-sided thickening of a world-line.



l / Bell's spaceship paradox.

along with the trailing rocket after the acceleration has been going on for a while. The x' axis is a line of simultaneity for \mathbf{o}' , and this axis intersects the leading ship's world-line at a point that \mathbf{o} considers to be *later* in time. Therefore \mathbf{o}' says that the leading ship has reached a higher speed than the trailing one. In \mathbf{o}' , the two ships' accelerations are unequal.

We can also see directly from the spacetime diagram that whereas length L_1 is 4 units as measured by an observer initially at rest relative to the thread, L_2 is about 5 units as measured by \mathbf{o}' , who is at rest relative to the trailing end of the thread at a later time. Since L_2 is greater than the unstressed length L_1 , the thread is under tension.

Figure 1/3 is more in the spirit of Bell's analysis. In frame \mathbf{o} , the thread has initial, unstressed length L . If the thread had been attached only to the leading ship, then it would have trailed behind it, unstressed, with Lorentz-contracted length L/γ . Since its actual length according to \mathbf{o} is still L , it has been stretched relative to its unstressed length.

This paradox relates to the difficult philosophical question of whether the time dilation and length contractions predicted by relativity are "real." This depends, of course, on what one means by "real." They are frame-dependent, i.e., observers in different frames of reference disagree about them. But this doesn't tell us much about their reality, since velocities are frame-dependent in Newtonian mechanics, but nobody worries about whether velocities are real. Bell took his colleagues' wrong answers as evidence that their intuitions had been misguided by the standard way of approaching this question of the reality of Lorentz contractions.

This treatment has one wart on it, which is that we judged the distance between the two ships in a frame of reference instantaneously comoving with the trailing ship, but this is slightly different from the length as determined in the leading ship's frame. One way to remove this wart is to note that the fractional discrepancy $\Delta L_1/L_1$ is of order v^3 , which is of a lower order than the strain in the thread, which is of order v^2 . To carry out this type of error estimation rigorously would however be cumbersome. A more elegant and rigorous approach is given in section 9.5.5 on p. 206, where we use fancier techniques to show that the motion shown in figure k is the unique motion that allows every portion of the string to move without strain.

This presentation includes ideas contributed by physicsforums users tiny-tim and PeterDonis.

3.9.3 Deja vu, jamais vu

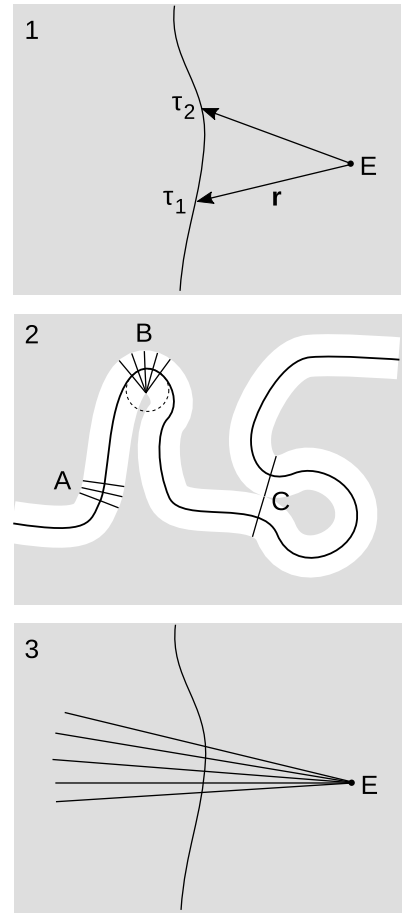
Deja vu all over again

In example 17 on p. 34, we saw that when an observer is accelerated, she may consider an event to be simultaneous with her more than once. That is, given a smooth, timelike world-line $\mathbf{r}(\tau)$ parametrized by proper time, and an event E , which we take to be the origin of our coordinate system, there may be more than one time at which \mathbf{r} is orthogonal to the velocity vector \mathbf{v} (figure m/1). As remarked earlier, this is just a problem with applying a particular arbitrary labeling convention to a certain example — not an earthshaking crisis in physics. Nevertheless it is of some intrinsic geometric interest to characterize the circumstances under which it can happen. We would like to place some kind of bound on how much acceleration is needed and how distant E must be.

As a warm-up, consider the analogous problem in Euclidean space, m/2. Here we have the notion of a tubular neighborhood, which is the greatest thickening of a curve W such that no point in it lies on two different normals. The tubular neighborhood has a radius r , which is the greatest possible radius of a non-self-intersecting piece of rope whose central axis coincides with W . Normally, as in region A, the rope doesn't intersect itself. There are two qualitatively different reasons why the rope could self-intersect. One is local: the radius of curvature of W is too small, as at B, where W coincides with a circle of radius r . The other is global: two points that are far apart as measured along W could be close together in the ambient Euclidean space, as at point C.

If we carry over these ideas to Minkowski space, then the local case, m/3, is easy to analyze using the techniques we have developed. The analog of the radius of curvature is the inverse of the proper acceleration, which suggests that we should be able to get a bound on the radius of the tubular neighborhood in terms of the acceleration. Define $f(\tau) = \mathbf{r} \cdot \mathbf{v}$. At a given point on W , f is minus the Minkowski time coordinate that an observer whose world-line is W would assign, at that instant, to E . The condition for the type of self-intersection we're discussing is that both f and its derivative with respect to proper time f' vanish at the same point on W . Differentiating f using the product rule, we find $f' = \mathbf{v} \cdot \mathbf{v} + \mathbf{r} \cdot \mathbf{a} = 1 + \mathbf{r} \cdot \mathbf{a}$ (in the $+-$ signature), so that $\mathbf{r} \cdot \mathbf{a} = -1$.

We now make use of the fact that both \mathbf{a} and \mathbf{r} are orthogonal to W — the former as a general kinematic fact, and the latter because $f = 0$. This means that they lie in the plane perpendicular to W . The geometry of this plane is Euclidean, so we can apply the Euclidean inequality $|\mathbf{a} \cdot \mathbf{r}| \leq |\mathbf{a}| |\mathbf{r}|$, where the bars on the left denote the absolute value and the ones on the right the magnitudes of the vectors. We therefore have $|\mathbf{a}| |\mathbf{r}| \geq 1$. Since \mathbf{r} is



m / Deja vu.

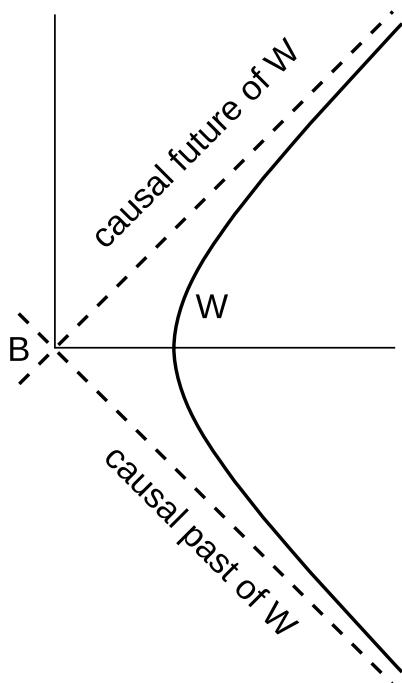
orthogonal to W , we can interpret it as the proper distance between E and W . The magnitude of \mathbf{a} is the proper acceleration. Converting to units with $c \neq 1$, we have an exact bound of the form (proper distance)(proper acceleration) $\geq c^2$. In ordinary units c is large, so in this sense E must be distant, and the acceleration large. This explains why we never encounter such a problem in nonrelativistic physics.

That was never now

So far we have characterized the circumstances under which simultaneity can fail to be unique. Simultaneity can also fail to exist. For example, in the same notation, take W to be the constant-acceleration motion described in example 4, p. 61, and let E be the event $(-1, 0)$. Then it can easily be shown (problem 17) that $f(\tau)$ is always positive, so an observer moving along W will always consider E to be in her past, never her future. No time exists for her such that she considers E to be “now.” The function f comes to a maximum somewhere but never crosses zero.

There will always be some neighborhood of W within which we are protected against nonexistence of simultaneity. To determine the radius of this neighborhood, we consider an event B that lies on the boundary of the neighborhood, and define f in terms of B rather than E . Then $f(\tau) = 0$ for some τ , but f does not cross over zero, so that either $f \leq 0$ everywhere or $f \geq 0$ everywhere. At the place where $f(\tau) = 0$, we also have $f'(\tau) = 0$, and the rest of the analysis is the same as before. Therefore the radius of the tubular neighborhood determined in that example defines a radius within which simultaneity has both existence and uniqueness.

The interpretation of such a boundary point B is a little funny. Figure n recapitulates the motion described in example 4. For this motion, the only point like B is the one labeled $(0, 0)$ in the Minkowski coordinates used in that derivation. Is there really anything special about this point, or is it just a random point that we happened to choose as the origin of our coordinate system? An observer moving along this W does not believe that any point in spacetime accessible to her has any special properties. She has always been accelerating and always will be, so no event she can observe or affect can be distinguished from the other events that she could have observed or affected in the same way at any earlier or later time. But we can easily show that B *is* special by giving a description of it without reference to any coordinates. Let W 's causal future be the set of all events that lie in the future light cone of some event on W , and similarly for W 's causal past. The boundaries of these two sets are W 's past and future event horizons, and these horizons coincide at only one event, which is B . This seems paradoxical, but our observer can neither observe nor affect B , so there is no contradiction.



n / The boundary point B really is special.

Problems

1 Fred buys a ticket on a spaceship that will accelerate to an ultrarelativistic speed v such that $c - v$ is only 6 m/s. Fred was on the track team in high school, so he knows he can run about 8 m/s. Once the ship is up to speed, Fred plans to run in the forward direction, thereby becoming the first human to exceed the speed of light. Other than the possible lack of gravity to allow running, what is wrong with Fred's plan?

2 (a) In the equation $v_c = (v_1 + v_2)/(1 + v_1 v_2)$ for combination of velocities, interpret the case where one of the velocities (but not the other) equals the speed of light. (b) Interpret the case where the denominator goes to zero. (c) Use the geometric series to rewrite the factor $1/(1 + v_1 v_2)$, and then expand the expression for v_c as a series in v_1 and v_2 , retaining terms up to third order in velocity. How does this relate to the correspondence principle?

3 Determine which of the identities in section 3.6 need to be modified in order to be valid in units with $c \neq 1$, and describe how they should be modified.

4 The Large Hadron Collider accelerates counterrotating beams of protons and collides them head-on. The beam energy has been gradually increased, and the accelerator is designed to reach a maximum energy of 14 TeV, corresponding to a rapidity of 10.3. (a) Find the velocity of the beam. (b) In any collision, the kinetic energy available to do something inelastic (smash up your car, produce nuclear reactions, ...) is the energy in the center of mass frame; in any other frame, there is initial kinetic energy that must also be present in the final state due to conservation of momentum. Suppose that a particular proton in the LHC beam never undergoes a collision with a proton from the opposite beam, and instead is wasted by being dumped into a beamstop. Let's say that this collision is with a proton in a hydrogen atom left behind by someone's fingerprint. Find the velocities of the two protons in their common center of mass frame.

5 Each GPS satellite is in an orbit with a radius of 26,600 km, with an orbital period of half a sidereal day, giving it a velocity of 3.88 km/s. The atomic clock aboard such a satellite is tuned to 10.22999999543 MHz, which is chosen so that when the satellite is directly overhead, the effect of time dilation (transverse Doppler shift), combined with a general-relativistic effect due to gravity, results in a frequency of exactly 10.23 MHz. (GPS started out as a military project, and legend has it that the top brass, suspicious of the crazy relativity stuff, demanded that the satellites be equipped with a software switch to turn off the correction, just in case the physicists were wrong.) There are oscillations superimposed onto these static effects due to the longitudinal Doppler shifts as the satellites approach and recede from a given observer on the ground.

(a) Calculate the maximum Doppler-shifted frequency for a hypothetical observer in outer space who is being directly approached by the satellite in its orbit. (b) In reality, the greatest possible longitudinal component of the velocity is considerably smaller than this due to the geometry. Use the size of the earth to determine this velocity and the corresponding maximum frequency.

6 Verify directly, using the geometry of figure b/2 on p. 55 that for $v = 3/5$, the Doppler shift factor is $D = 2$. (Do not simply plug $v = 3/5$ into the formula $D = \sqrt{(1+v)/(1-v)}$.)

7 Generalize the numerical calculation of problem 6 to prove the general result $D = \sqrt{(1+v)/(1-v)}$.

8 Expand the relativistic equation for the longitudinal Doppler shift of light $D(v)$ in a Taylor series, and find the first two nonvanishing terms. Show that these two terms agree with the nonrelativistic expression, so that any relativistic effect is of higher order in v .

9 Prove, as claimed on p. 64, that we must have $\mathbf{a} \cdot \mathbf{v} = 0$ if the velocity four-vector is to remain properly normalized.

10 Example 4 on p. 61 described the motion of an object having constant proper acceleration a , the world-line being $t = \frac{1}{a} \sinh a\tau$ and $x = \frac{1}{a} \cosh a\tau$ in a particular observer's Minkowski coordinates. (a) Prove the following results for γ and for the (three-)velocity and (three-)acceleration measured by this observer.

$$\gamma = \cosh a\tau$$

$$v = \tanh a\tau$$

$$\text{acceleration} = a \cosh^{-3} a\tau$$

Do the calculations simply by taking the first and second derivatives of position with respect to time. You will find the following facts helpful:

$$1 - \tanh^2 = \cosh^{-2}$$

$$\frac{d}{dx} \tanh x = \cosh^{-2} x$$

(b) Interpret the results in the limit of large τ .

11 Example 4 on p. 61 described the motion of an object having constant proper acceleration a , the world-line being $t = \frac{1}{a} \sinh a\tau$ and $x = \frac{1}{a} \cosh a\tau$ in a particular observer's Minkowski coordinates. Find the corresponding velocity and acceleration four-vectors.

12 Starting from the results of problem 11, repeat problem 10a using the techniques of section 3.7 on p. 66. You will find it helpful to know that $1 - \tanh^2 = \cosh^{-2}$.

13 Let \mathbf{v} be a future-directed, properly normalized velocity vector. Compare the value of $\mathbf{v} \cdot \mathbf{v}$ in the $+- --$ signature used in this book with its value in the signature $-+++$.

14 (a) Prove the relation $d\gamma/dt = \gamma^3 \mathbf{a}_0 \cdot \mathbf{v}_0$ given on p. 67, in the special case where the motion is linear. (b) Generalize the result to $3 + 1$ dimensions.

15 Derive the identity $\mathbf{a}_0 = \frac{1}{(\mathbf{o} \cdot \mathbf{v})^2} [P_0 \mathbf{a} - (\mathbf{o} \cdot \mathbf{a}) \mathbf{v}_0]$ on p. 67.

16 Recapitulating the geometry in figure m on p. 73, let W be a smooth, timelike world line, E an event not on W , and \mathbf{r} the vector from E to a point on W , parametrized by proper time τ . Define the proper distance ℓ between E and a point on W as $\ell^2 = -(P_{\mathbf{v}} \mathbf{r})^2$, where the square indicates an inner product of the vector with itself, and the minus sign is because we use the $+- --$ signature. Show that $d(\ell^2)/d\tau = 2(\mathbf{r} \cdot \mathbf{v})(\mathbf{r} \cdot \mathbf{a})(\mathbf{v} \cdot \mathbf{v})$, where the final factor is just a signature-dependent sign. Does this make sense when W is inertial? Give an example where the derivative vanishes because the first factor is zero, and another example where the second factor is the one that vanishes (but $\mathbf{a} \neq 0$).

17 Consider an observer O moving along a world-line W with the constant-acceleration motion defined in example 4, p. 61. In section 3.9.3, p. 73, we gave the coordinates of a certain event E that was never “now” as described by our observer. The purpose of this problem is to analyze this in a more elegant and coordinate-invariant way. Let P be a point on W , let B be the event described in section 3.9.3, and let $\mathbf{x} = \overrightarrow{BP}$, $\mathbf{h} = \overrightarrow{BE}$, and $\mathbf{r} = \overrightarrow{EP}$. (a) Show that W , which was originally described in a certain set of coordinates, can instead be defined by the fact that $\mathbf{x} \cdot \mathbf{v} = 0$ for every point on W . (b) Show that if \mathbf{h} is timelike, then $\mathbf{r} \cdot \mathbf{v}$ is never zero.

Chapter 4

Dynamics

4.1 Ultrarelativistic particles

A typical 22-caliber rifle shoots a bullet with a mass of about 3 g at a speed of about 400 m/s. Now consider the firing of such a rifle as seen through an ultra-powerful telescope by an alien in a distant galaxy. We happen to be firing in the direction away from the alien, who gets a view from over our shoulder. Since the universe is expanding, our two galaxies are receding from each other. In the alien's frame, our own galaxy is the one that is moving — let's say at¹ $c - (200 \text{ m/s})$. If the two velocities simply added, the bullet would be moving at $c + (200 \text{ m/s})$. But velocities don't simply add and subtract relativistically, and applying the correct equation for relativistic combination of velocities, we find that in the alien's frame, the bullet flies at only $c - (199.9995 \text{ m/s})$. That is, according to the alien, the energy in the gunpowder only succeeded in accelerating the bullet by 0.0005 m/s! If we insisted on believing in $K = (1/2)mv^2$, this would clearly violate conservation of energy in the alien's frame of reference. It appears that kinetic energy must not only rise faster than v^2 as v approaches c , it must blow up to infinity. This gives a dynamical explanation for why no material object can ever reach or exceed c , as we have already inferred on purely kinematical grounds.

To the alien, both our galaxy and the bullet are ultrarelativistic objects, i.e., objects moving at nearly c . A good way of thinking about an ultrarelativistic particle is that it's a particle with a very small mass. For example, the subatomic particle called the neutrino has a very small mass, thousands of times smaller than that of the electron. Neutrinos are emitted in radioactive decay, and because the neutrino's mass is so small, the amount of energy available in these decays is always enough to accelerate it to very close to c . Nobody has ever succeeded in observing a neutrino that was *not* ultrarelativistic. When a particle's mass is very small, the mass becomes difficult to measure. For almost 70 years after the neutrino was discovered, its mass was thought to be zero. Similarly, we currently believe that a ray of light has no mass, but it is always possible that its mass will be found to be nonzero at some point

¹In reality when two velocities move at relativistic speeds compared with one another, they are separated by a cosmological distance, and special relativity does not actually allow us to construct frames of reference this large.

in the future. A ray of light can be modeled as an ultrarelativistic particle.

Let's compare ultrarelativistic particles with train cars. A single car with kinetic energy E has different properties than a train of two cars each with kinetic energy $E/2$. The single car has half the mass and a speed that is greater by a factor of $\sqrt{2}$. But the same is not true for ultrarelativistic particles. Since an idealized ultrarelativistic particle has a mass too small to be detectable in any experiment, we can't detect the difference between m and $2m$. Furthermore, ultrarelativistic particles move at close to c , so there is no observable difference in speed. Thus we expect that a single ultrarelativistic particle with energy E compared with two such particles, each with energy $E/2$, should have all the same properties as measured by a mechanical detector.

An idealized zero-mass particle also has no frame in which it can be at rest. It always travels at c , and no matter how fast we chase after it, we can never catch up. We can, however, observe it in different frames of reference, and we will find that its energy is different. For example, distant galaxies are receding from us at substantial fractions of c , and when we observe them through a telescope, they appear very dim not just because they are very far away but also because their light has less energy in our frame than in a frame at rest relative to the source. This effect must be such that changing frames of reference according to a specific Lorentz transformation always changes the energy of the particle by a fixed factor, regardless of the particle's original energy; for if not, then the effect of a Lorentz transformation on a single particle of energy E would be different from its effect on two particles of energy $E/2$.

How does this energy-shift factor depend on the velocity v of the Lorentz transformation? Here it becomes nicer to work in terms of the variable D . Let's write $f(D)$ for the energy-shift factor that results from a given Lorentz transformation. Since a Lorentz transformation D_1 followed by a second transformation D_2 is equivalent to a single transformation by $D_1 D_2$, we must have $f(D_1 D_2) = f(D_1) f(D_2)$. This tightly constrains the form of the function f ; it must be something like $f(D) = D^n$, where n is a constant. The interpretation of n is that under a Lorentz transformation corresponding to 1% of c , energies of ultrarelativistic particles change by about $n\%$ (making the approximation that $v = .01$ gives $D \approx 1.01$). In his original 1905 paper on special relativity, Einstein used Maxwell's equations and the Lorentz transformation to show that for a light wave $n = 1$, and we will prove on p. 88 that this holds for any ultrarelativistic object. He wrote, "It is remarkable that the energy and the frequency ... vary with the state of motion of the observer in accordance with the same law." He was presumably interested in this fact because 1905 was also the year in which he published his paper on the photoelectric effect, which formed the

foundations of quantum mechanics. An axiom of quantum mechanics is that the energy and frequency of any particle are related by $E = hf$, and if E and f hadn't transformed in the same way relativistically, then quantum mechanics would have been incompatible with relativity.

If we assume that certain objects, such as light rays, are truly massless, rather than just having masses too small to be detectable, then their D doesn't have any finite value, but we can still find how the energy differs according to different observers by finding the D of the Lorentz transformation between the two observers' frames of reference.

An astronomical energy shift

Example 1

▷ For quantum-mechanical reasons, a hydrogen atom can only exist in states with certain specific energies. By conservation of energy, the atom can therefore only absorb or emit light that has an energy equal to the difference between two such atomic energies. The outer atmosphere of a star is mostly made of monoatomic hydrogen, and one of the energies that a hydrogen atom can absorb or emit is 3.0276×10^{-19} J. When we observe light from stars in the Andromeda Galaxy, it has an energy of 3.0306×10^{-19} J. If this is assumed to be due entirely to the motion of the Milky Way and Andromeda Galaxy relative to one another, along the line connecting them, find the direction and magnitude of this velocity.

▷ The energy is shifted upward, which means that the Andromeda Galaxy is moving toward us. (Galaxies at cosmological distances are always observed to be receding from one another, but this doesn't necessarily hold for galaxies as close as these.) Relating the energy shift to the velocity, we have

$$\frac{E'}{E} = D = \sqrt{(1 + v)/(1 - v)}.$$

Since the shift is only about one part per thousand, the velocity is small compared to c — or small compared to 1 in units where $c = 1$. Therefore we can employ the low-velocity approximation $D \approx 1 + v$, which gives

$$v \approx D - 1 = \frac{E'}{E} - 1 = -1.0 \times 10^{-3}.$$

The negative sign confirms that the source is approaching rather than receding. This is in units where $c = 1$. Converting to SI units, where $c \neq 1$, we have $v = (-1.0 \times 10^{-3})c = -300$ km/s. Although the Andromeda Galaxy's tangential motion is not accurately known, it is considered likely that it will collide with the Milky Way in a few billion years.

4.2 $E=mc^2$

We now know the relativistic expression for kinetic energy in the limiting case of an ultrarelativistic particle: its energy is proportional to the “stretch factor” D of the Lorentz transformation. What about intermediate cases, like $v = c/2$?

a / The match is lit inside the bell jar. It burns, and energy escapes from the jar in the form of light. After it stops burning, all the same atoms are still in the jar: none have entered or escaped. The figure shows the outcome expected before relativity, which was that the mass measured on the balance would remain exactly the same. This is not what happens in reality.



When we are forced to tinker with a time-honored theory, our first instinct should always be to tinker as conservatively as possible. Although we’ve been forced to admit that kinetic energy doesn’t vary as $v^2/2$ at relativistic speeds, the next most conservative thing we could do would be to assume that the *only* change necessary is to replace the factor of $v^2/2$ in the nonrelativistic expression for kinetic energy with some other function, which would have to act like D or $1/D$ for $v \rightarrow \pm c$. I suspect that this is what Einstein thought when he completed his original paper on relativity in 1905, because it wasn’t until later that year that he published a second paper showing that this still wasn’t enough of a change to produce a working theory. We now know that there is something more that needs to be changed about prerelativistic physics, and this is the assumption that mass is only a property of material particles such as atoms (figure a). Call this the “atoms-only hypothesis.”

Now that we know the correct relativistic way of finding the energy of a ray of light, it turns out that we can use that to find what we were originally seeking, which was the energy of a material object. The following discussion closely follows Einstein’s.

Suppose that a material object O of mass m_o , initially at rest in a certain frame A , emits two rays of light (or any other kind of ultrarelativistic particles), each with energy $E/2$. By conservation of energy, the object must have lost an amount of energy equal to E . By symmetry, O remains at rest.

We now switch to a different frame of reference B moving at some arbitrary speed corresponding to a stretch factor D . The change of frames means that we’re chasing one ray, so that its energy is scaled down to $(E/2)D^{-1}$, while running away from the other, whose energy gets boosted to $(E/2)D$. In frame B , as in A , O retains the

same speed after emission of the light. But observers in frames A and B disagree on how much energy O has lost, the discrepancy being

$$E \left[\frac{1}{2}(D + D^{-1}) - 1 \right].$$

This can be rewritten using identity [2] from section 3.6 as

$$E(\gamma - 1).$$

Let's consider the case where B's velocity relative to A is small. Using the approximation $\gamma \approx 1 + v^2/2$, our result is approximately

$$\frac{1}{2}Ev^2,$$

neglecting terms of order v^4 and higher. The interpretation is that when O reduced its energy by E in order to make the light rays, it reduced its *mass* from m_o to $m_o - m$, where $m = E$. Inserting the necessary factor of c^2 to make this valid in units where $c \neq 1$, we have Einstein's famous

$$E = mc^2.$$

This derivation entailed both an approximation and some hidden assumptions. These issues are explored more thoroughly in section 4.4 on p. 98 and in ch. 9 on p. 175. The result turns out to be valid for any isolated body.

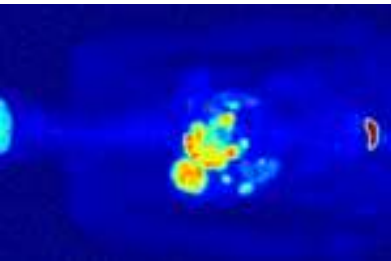
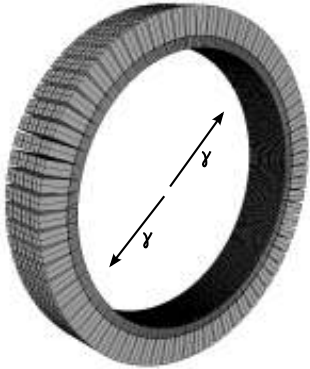
We find that mass is not simply a built-in property of the particles that make up an object, with the object's mass being the sum of the masses of its particles. Rather, mass and energy are equivalent, so that if the experiment of figure a is carried out with a sufficiently precise balance, the reading will drop because of the mass equivalent of the energy emitted as light.

The equation $E = mc^2$ tells us how much energy is equivalent to how much mass: the conversion factor is the square of the speed of light, c . Since c a big number, you get a really really big number when you multiply it by itself to get c^2 . This means that even a small amount of mass is equivalent to a very large amount of energy. Conversely, an ordinary amount of energy corresponds to an extremely small mass, and this is why nobody detected the non-null result of experiments like the one in figure a hundreds of years ago.

The big event here is mass-energy equivalence, but we can also harvest a result for the energy of a material particle moving at a certain speed. We have $m(\gamma - 1)$ for the difference between O's energy in frame B and its energy when it is at rest, i.e., its kinetic energy. But since mass and energy are equivalent, we assign O an energy m when it is at rest. The result is that the energy is

$$E = m\gamma$$

(or $m\gamma c^2$ in units with $c \neq 1$).



b / Top: A PET scanner. Middle: Each positron annihilates with an electron, producing two gamma-rays that fly off back-to-back. When two gamma rays are observed simultaneously in the ring of detectors, they are assumed to come from the same annihilation event, and the point at which they were emitted must lie on the line connecting the two detectors. Bottom: A scan of a person's torso. The body has concentrated the radioactive tracer around the stomach, indicating an abnormal medical condition.

Electron-positron annihilation

Example 2

Natural radioactivity in the earth produces positrons, which are like electrons but have the opposite charge. A form of antimatter, positrons annihilate with electrons to produce gamma rays, a form of high-frequency light. Such a process would have been considered impossible before Einstein, because conservation of mass and energy were believed to be separate principles, and this process eliminates 100% of the original mass. The amount of energy produced by annihilating 1 kg of matter with 1 kg of antimatter is

$$\begin{aligned} E &= mc^2 \\ &= (2 \text{ kg}) (3.0 \times 10^8 \text{ m/s})^2 \\ &= 2 \times 10^{17} \text{ J}, \end{aligned}$$

which is on the same order of magnitude as a day's energy consumption for the entire world's population!

Positron annihilation forms the basis for the medical imaging technique called a PET (positron emission tomography) scan, in which a positron-emitting chemical is injected into the patient and mapped by the emission of gamma rays from the parts of the body where it accumulates.

A rusting nail

Example 3

▷ An iron nail is left in a cup of water until it turns entirely to rust. The energy released is about 0.5 MJ. In theory, would a sufficiently precise scale register a change in mass? If so, how much?

▷ The energy will appear as heat, which will be lost to the environment. The total mass-energy of the cup, water, and iron will indeed be lessened by 0.5 MJ. (If it had been perfectly insulated, there would have been no change, since the heat energy would have been trapped in the cup.) The speed of light is $c = 3 \times 10^8$ meters per second, so converting to mass units, we have

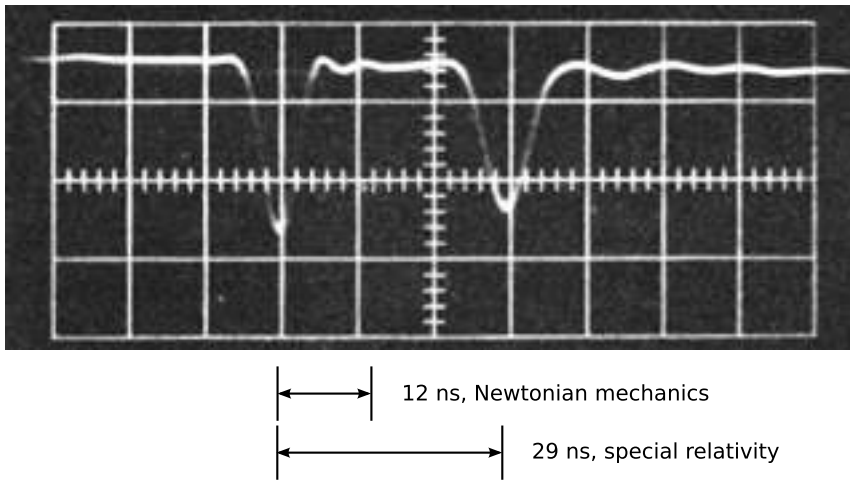
$$\begin{aligned} m &= \frac{E}{c^2} \\ &= \frac{0.5 \times 10^6 \text{ J}}{(3 \times 10^8 \text{ m/s})^2} \\ &= 6 \times 10^{-12} \text{ kilograms}. \end{aligned}$$

The change in mass is too small to measure with any practical technique. This is because the square of the speed of light is such a large number.

Relativistic kinetic energy

Example 4

By about 1930, particle accelerators had progressed to the point at which relativistic effects were routinely taken into account. In 1964, W. Bertozzi did a special-purpose experiment to test the predictions of relativity using an electron accelerator. The results were discussed in less detail in example 2 on p. 58, at which point we had not yet seen the relativistic equation for kinetic energy. Electrons were accelerated through a static electric potential difference V to a variety of kinetic energies $K = eV$, and their velocities inferred by measuring their time of flight through a beamline of length $\ell = 8.4$ m. Electrical pulses were recorded on an oscilloscope at the beginning and end of the time of flight t . The energies were confirmed by calorimetry. Figure c shows a sample photograph of an oscilloscope trace at $V = 1.5$ MeV.



c / Example 4. Each horizontal division is 9.8 ns.

The prediction of Newtonian physics is as follows.

$$eV = (1/2)mv^2$$

$$v/c = 2.4$$

$$t = 12 \text{ ns}$$

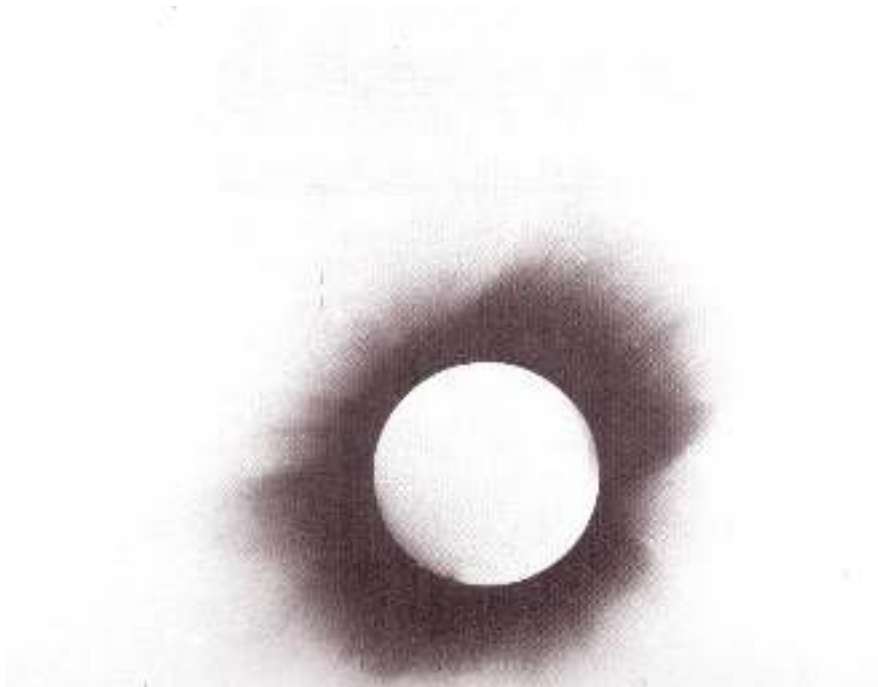
According to special relativity, we have:

$$eV = m(\gamma - 1)c^2$$

$$\frac{v}{c} = \sqrt{1 - \left(1 + \frac{eV}{mc^2}\right)^{-2}} = 0.97$$

$$t = 29 \text{ ns}$$

The results contradict the Newtonian prediction and are consistent with special relativity. According to Newton, this amount of energy should have accelerated the electrons to several times the speed of light. In reality, we see a clear demonstration of the nature of c as a limiting velocity.



LIGHTS ALL ASKEW IN THE HEAVENS

Men of Science More or Less
Agog Over Results of Eclipse
Observations.

EINSTEIN THEORY TRIUMPHS

Stars Not Where They Seemed
or Were Calculated to be,
but Nobody Need Worry.

A BOOK FOR 12 WISE MEN

No More in All the World Could
Comprehend It, Said Einstein When
His Daring Publishers Accepted It.

Gravity bending light

Example 5

Gravity is a universal attraction between things that have mass, and since the energy in a beam of light is equivalent to some very small amount of mass, light should be affected by gravity, although the effect should be very small. The first experimental confirmation of relativity came in 1919 when stars next to the sun during a solar eclipse were observed to have shifted a little from their ordinary position. (If there was no eclipse, the glare of the sun would prevent the stars from being observed.) Starlight had been deflected by the sun's gravity. The figure is a photographic negative, so the circle that appears bright is actually the dark face of the moon, and the dark area is really the bright corona of the sun. The stars, marked by lines above and below then, appeared at positions slightly different than their normal ones.

Keep in mind that these arguments are very rough and qualitative, and it is *not* possible to produce a relativistic theory of gravity simply by taking $E = mc^2$ and combining it with Newton's law of gravity. After all, this law doesn't refer to time at all: it predicts that gravitational forces propagate instantaneously. We know this can't be consistent with relativity, which forbids cause and effect from propagating at any speed greater than c . To produce a relativistic theory of gravity, we need general relativity.

Similar reasoning suggests that there may be stars — black holes — so dense that their gravity can prevent light from leaving. Such stars have been detected, and their properties seem so far to be described correctly by general relativity.

4.3 Relativistic momentum

4.3.1 The energy-momentum vector

Newtonian mechanics has two different measures of motion, kinetic energy and momentum, and the relationship between them is nonlinear. Doubling your car's momentum quadruples its kinetic energy.

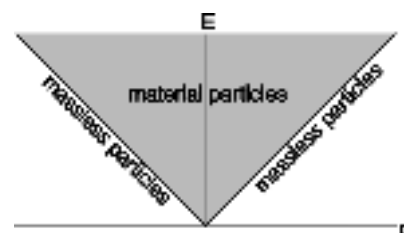
But nonrelativistic mechanics can't handle massless particles, which are always ultrarelativistic. We saw in section 4.1 that ultrarelativistic particles are “generic,” in the sense that they have no individual mechanical properties other than an energy and a direction of motion. Therefore the relationship between kinetic energy and momentum must be *linear* for ultrarelativistic particles. For example, doubling the amplitude of an electromagnetic wave quadruples both its energy density, which depends on \mathbf{E}^2 and \mathbf{B}^2 , and its momentum density, which goes like $\mathbf{E} \times \mathbf{B}$.

How can we make sense of these energy-momentum relationships, which seem to take on two completely different forms in the limiting cases of very low and very high velocities?

The first step is realize that since mass and energy are equivalent, we will get more of an apples-to-apples comparison if we stop talking about a material object's *kinetic* energy and consider instead its *total* energy E , which includes a contribution from its mass.

Figure d is a graph of energy versus momentum. In this representation, massless particles, which have $E \propto |p|$, lie on two diagonal lines that connect at the origin. If we like, we can pick units such that the slopes of these lines are plus and minus one. Material particles lie above these lines. For example, a car sitting in a parking lot has $p = 0$ and $E = m$.

Now what happens to such a graph when we change to a different frame or reference that is in motion relative to the original frame? A massless particle still has to act like a massless particle, so the diagonals are simply stretched or contracted along their own lengths. A transformation that always takes a line to a line is a linear transformation, and if the transformation between different frames of reference preserves the linearity of the lines $p = E$ and $p = -E$, then it's natural to suspect that it is actually some kind of linear transformation. In fact the transformation must be linear, because conservation of energy and momentum involve addition, and we need these laws to be valid in all frames of reference. But now by the same reasoning as in subsection 1.3.1 on p. 22, the transformation must be area-preserving. We then have the same three cases to consider as in figure j on p. 16. The “Galilean” version is ruled out because it would imply that particles keep the same energy when we change frames. (This is what would happen if c were infinite, so that the mass-equivalent E/c^2 of a given energy was zero,



d / In the p - E plane, massless particles lie on the two diagonals, while particles with mass lie to the right.

and therefore E would be interpreted purely as the mass.) Nor can the “rotational” version be right, because it doesn’t preserve the $E = |p|$ diagonals. We are left with the third case, which establishes the following aesthetically appealing fact:

Energy-momentum is a four-vector

Let an isolated object have momentum and mass-energy p and E . Then the p - E plane transforms according to exactly the same kind of Lorentz transformation as the x - t plane. That is, (E, p_x, p_y, p_z) is a four-dimensional vector just like (t, x, y, z) .

This is a highly desirable result. If it were not true, it would be like having to learn different mathematical rules for different kinds of three-vectors in Newtonian mechanics.

The only remaining issue to settle is whether the choice of units that gives invariant 45-degree diagonals in the x - t plane is the same as the choice of units that gives such diagonals in the p - E plane. That is, we need to establish that the c that applies to x and t is equal to the c' needed for p and E , i.e., that the velocity scales of the two graphs are matched up. This is true because in the Newtonian limit, the total mass-energy E is essentially just the particle’s mass, and then $p/E \approx p/m \approx v$. This establishes that the velocity scales are matched at small velocities, which implies that they coincide for all velocities, since a large velocity, even one approaching c , can be built up from many small increments. (This also establishes that the exponent n defined on p. 80 equals 1 as claimed.)

Suppose that a particle is at rest. Then it has $p = 0$ and mass-energy E equal to its mass m . Therefore the inner product of its (E, p) four-vector with itself equals m^2 . In other words, the “magnitude” of the energy-momentum four-vector is simply equal to the particle’s mass. If we transform into a different frame of reference, in which $p \neq 0$, the inner product stays the same. In symbols,

$$m^2 = E^2 - p^2,$$

or, in units with $c \neq 1$,

$$(mc^2)^2 = E^2 - (pc)^2.$$

We take this as the relativistic definition of mass. Since the definition is an inner product, which is a scalar, it is the same in all frames of reference. (Some older books use an obsolete convention of referring to $m\gamma$ as “mass” and m as “rest mass.”)

self-check A

Interpret the equation $m^2 = E^2 - p^2$ in the case where $m = 0$. ▷

Answer, p. ??

A high-precision test of this fundamental relativistic relationship was carried out by Meyer *et al.* in 1963 by studying the motion of electrons in static electric and magnetic fields. They define the quantity

$$Y^2 = \frac{E^2}{m^2 + p^2},$$

which according to special relativity should equal 1. Their results, tabulated in the sidebar, show excellent agreement with theory.

Results from Meyer et al., 1963

v	Y
0.9870	1.0002(5)
0.9881	1.0012(5)
0.9900	0.9998(5)

Mass of two light waves

Example 6

Let the momentum of a certain light wave be $(p_t, p_x) = (E, E)$, and let another such wave have momentum $(E, -E)$. The total momentum is $(2E, 0)$. Thus this pair of massless particles has a collective mass of $2E$. This is an example of the non-additivity of relativistic mass.

4.3.2 Collision invariants

Example 6 shows that mass is not additive, nor it is a measure of the “quantity of matter.” More generally, suppose that we have a collision between two objects, which could be two cars or two nuclei in a particle accelerator. Conservation of (spatial) momentum dictates that not all the energy is available for smashing windshields or creating gamma rays. For example, a Martian watching a parking-lot fender-bender through a powerful telescope would say that both cars were going as fast as fighter jets, due to the rotation of the earth, but this doesn’t make the bang any louder. To avoid being misled by these frame-dependent distractions, we can concentrate only on quantities that are scalars. For a two-body collision, there are three such scalars that we can construct: \mathbf{p}_1^2 , \mathbf{p}_2^2 , and $\mathbf{p}_1 \cdot \mathbf{p}_2$. (The notation \mathbf{a}^2 is simply an abbreviation for $\mathbf{a} \cdot \mathbf{a}$.) These are known as the collision invariants. The first two of these are simply the squared masses of the individual particles.

Now consider the center of mass frame, i.e., the frame in which the total momentum has a zero spacelike part. In this frame, the total energy-momentum vector is of the form $(E_{\text{cm}}, 0)$, corresponding to a mass $M = E_{\text{cm}}$. All of this energy is available to make a bang. If we were colliding particles in an accelerator in order to produce new particles, this collision would be just barely enough energy to create a single particle of mass M , if the two incoming particles were annihilated in the process. This center of mass energy can be expressed in terms of the collision invariants as $M^2 = (\mathbf{p}_1 + \mathbf{p}_2)^2 = \mathbf{p}_1^2 + \mathbf{p}_2^2 + 2\mathbf{p}_1 \cdot \mathbf{p}_2 = m_1^2 + m_2^2 + 2\mathbf{p}_1 \cdot \mathbf{p}_2$. This is a nonlinear relationship, and the third collision invariant $\mathbf{p}_1 \cdot \mathbf{p}_2$ tells us how the nonlinearity plays out based on the relative directions of motion. The two momentum vectors are both timelike and future-directed, so by the reversed triangle equality (section 1.5, p. 36) we have $M \geq m_1 + m_2$.

4.3.3 Some examples involving momentum

Finding velocity given energy and momentum *Example 7*

▷ If we know that a particle has mass-energy E and momentum p (which also implies knowledge of its mass m), what is its velocity?

▷ In the particle's rest frame it has a world-line that points straight up on a spacetime diagram, and its momentum vector \mathbf{p} likewise points up in the $p - E$ plane. Since displacement vectors and momentum vectors transform according to the same rules, this parallelism will be maintained in other frames as well. Therefore in an arbitrarily chosen frame, the vector $\mathbf{p} = (E, p)$ lies along a line whose inverse slope $v = p/E$ gives the velocity.

As a check on our result, we look at its limiting behavior. In the Newtonian limit, the mass-energy E is nearly all due to the mass, so we have $v \approx p/m$, the Newtonian result. In the opposite limit of ultrarelativistic motion, with $E \gg m$, the definition of mass $m^2 = E^2 - p^2$ gives $E \approx |p|$, and we have $|v| \approx 1$, which is also correct.

Light rays don't interact *Example 8*

We observe that when two rays of light cross paths, they continue through one another without bouncing like material objects. This behavior follows directly from conservation of energy-momentum.

Any two vectors can be contained in a single plane, so we can choose our coordinates so that both rays have vanishing p_z . By choosing the state of motion of our coordinate system appropriately, we can also make $p_y = 0$, so that the collision takes place along a single line parallel to the x axis. Since only p_x is nonzero, we write it simply as p . In the resulting p - E plane, there are two possibilities: either the rays both lie along the same diagonal, or they lie along different diagonals. If they lie along the same diagonal, then there can't be a collision, because the two rays are both moving in the same direction at the same speed c , and the trailing one will never catch up with the leading one.

Now suppose they lie along different diagonals. We add their energy-momentum vectors to get their total energy-momentum, which will lie in the gray area of figure d. That is, a pair of light rays taken as a single system act sort of like a material object with a nonzero mass. By a Lorentz transformation, we can always find a frame in which this total energy-momentum vector lies along the E axis. This is a frame in which the momenta of the two rays cancel, and we have a symmetric head-on collision between two rays of equal energy. It is the "center-of-mass" frame, although neither object has any mass on an individual basis. For convenience, let's assume that the x - y - z coordinate system was chosen so that its origin was at rest in this frame.

Since the collision occurs along the x axis, by symmetry it is not

possible for the rays after the collision to depart from the x axis; for if they did, then there would be nothing to determine the orientation of the plane in which they emerged.² Therefore we are justified in continuing to use the same p_x - E plane to analyze the four-vectors of the rays after the collision.

Let each ray have energy E in the frame described above. Given this total energy-momentum vector, how can we cook up two energy-momentum vectors for the final state such that energy and momentum will have been conserved? Since there is zero total momentum, our only choice is two light rays, one with energy-momentum vector (E, E) and one with $(E, -E)$. But this is exactly the same as our initial state, except that we can arbitrarily choose the roles of the two rays to have been interchanged. Such an interchanging is only a matter of labeling, so there is no observable sense in which the rays have collided.³

Compton scattering

Example 9

Figure e/1 is a histogram of gamma rays emitted by a ^{137}Cs source and recorded by a NaI scintillation detector. This type of detector, unlike a Geiger-Muller counter, gives a pulse whose height is proportional to the energy of the radiation. About half the gamma rays do what we would like them to do in a detector: they deposit their full energy of 662 keV in the detector, resulting in a prominent peak in the histogram. The other half, however, interact through a process called Compton scattering, in which they collide with one of the electrons but emerge from the collision still retaining some of their energy, with which they may escape from the detector. The amount of energy deposited in the detector depends solely on the billiard-ball kinematics of the collision, and can be determined from conservation of energy-momentum based on the scattering angle. Forward scattering at 0 degrees is no interaction at all, and deposits no energy, while scattering

²In quantum mechanics, there is a loophole here. Quantum mechanics allows certain kinds of randomness, so that the symmetry can be broken by letting the outgoing rays be observed in a plane with some random orientation.

³There is a second loophole here, which is that a ray of light is actually a wave, and a wave has other properties besides energy and momentum. It has a wavelength, and some waves also have a property called polarization. As a mechanical analogy for polarization, consider a rope stretched taut. Side-to-side vibrations can propagate along the rope, and these vibrations can occur in any plane that coincides with the rope. The orientation of this plane is referred to as the polarization of the wave. Returning to the case of the colliding light rays, it is possible to have nontrivial collisions in the sense that the rays could affect one another's wavelengths and polarizations. Although this doesn't actually happen with non-quantum-mechanical light waves, it can happen with other types of waves; see, e.g., Hu et al., arxiv.org/abs/hep-ph/9502276, figure 2. The title of example 8 is only valid if a "ray" is taken to be something that lacks wave structure. The wave nature of light is not evident in everyday life from observations with apparatus such as flashlights, mirrors, and eyeglasses, so we expect the result to hold under those circumstances, and it does. E.g., flashlight beams do pass through one another without interacting.

e / 1. The Compton edge lies at the energy deposited by gamma rays that scatter at 180 degrees from an electron. 2. The collision in the lab frame. 3. The same collision in the center of mass frame.

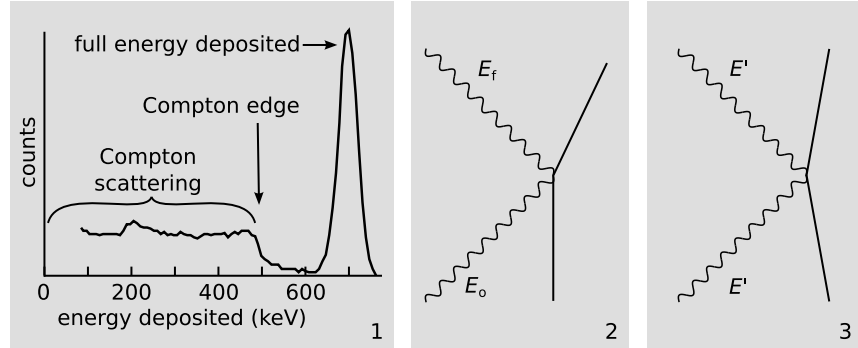


Figure e/2 shows the collision in the lab frame, where the electron is initially at rest. As is conventional in this type of diagram, the world-line of the photon is shown as a wiggly line; the wiggles are just a decoration, and the actual world-line consists of two line segments. The photon enters the detector with the full energy $E_0 = 662$ keV and leaves with a smaller energy E_f . The difference $E_0 - E_f$ is what the detector will measure, contributing a count to the Compton edge. In the lab frame, the total initial momentum vector is $\mathbf{p} = (E_0 + m, E_0)$, with the timelike component representing the total mass-energy. Because the photon is massless, its momentum $p_x = E_0$ is equal to its energy.

Let v be the velocity of the center-of-mass frame, e/3, relative to the lab frame. Using the result of example 7, we find $v = E_0 / (E_0 + m)$. To make the writing easier we define $\alpha = E_0 / m$, so that $v = \alpha / (1 + \alpha)$.

The transformation from the lab frame to the c.m. frame Doppler shifts the energy of the incident photon down to $E' = D(-v)E_0$. The collision reverses the spatial part of the photon's energy-momentum vector while leaving its energy the same. Transformation back into the lab frame gives $E_f = D(-v)E' = D(-v)^2 E_0 = E_0 / (1 + 2\alpha)$. (This can also be rewritten using the quantum-mechanical relation $E = hc/\lambda$ to give the compact form $\lambda_f - \lambda_0 = 2hc/m$.) The final result for the energy of the Compton edge is

$$\begin{aligned} E_0 - E_f &= \frac{E_0}{1 + 1/2\alpha} \\ &= 478 \text{ keV,} \end{aligned}$$

in good agreement with figure e/1.

Pair production requires matter *Example 10*
Example 2 on p. 84 discussed the annihilation of an electron and a positron into two gamma rays, which is an example of turning

matter into pure energy. An opposite example is pair production, a process in which a gamma ray disappears, and its energy goes into creating an electron and a positron.

Pair production cannot happen in a vacuum. For example, gamma rays from distant black holes can travel through empty space for thousands of years before being detected on earth, and they don't turn into electron-positron pairs before they can get here. Pair production can only happen in the presence of matter. When lead is used as shielding against gamma rays, one of the ways the gamma rays can be stopped in the lead is by undergoing pair production.

To see why pair production is forbidden in a vacuum, consider the process in the frame of reference in which the electron-positron pair has zero total momentum. In this frame, the gamma ray would have to have had zero momentum, but a gamma ray with zero momentum must have zero energy as well. This means that conservation of the momentum vector has been violated: the timelike component of the momentum is the mass-energy, and it has increased from 0 in the initial state to at least $2mc^2$ in the final state.

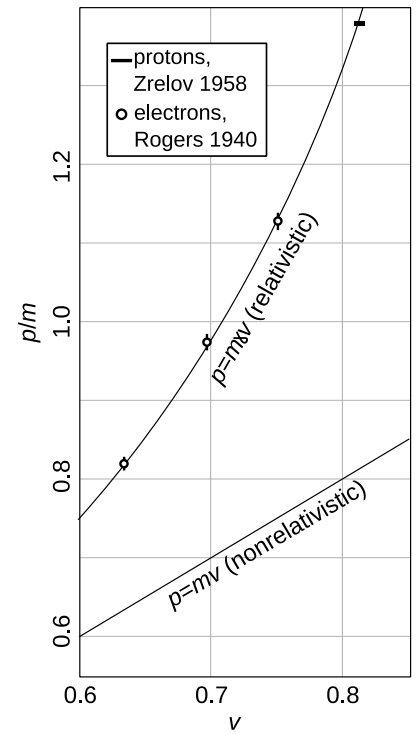
4.3.4 Massless particles travel at c

Massless particles always travel at $c(=1)$. For suppose that a massless particle had $|v| < 1$ in the frame of some observer. Then some other observer could be at rest relative to the particle. In such a frame, the particle's momentum p is zero by symmetry, since there is no preferred direction for it. Then $E^2 = p^2 + m^2$ is zero as well, so the particle's entire energy-momentum vector is zero. But a vector that vanishes in one frame also vanishes in every other frame. That means we're talking about a particle that can't undergo scattering, emission, or absorption, and is therefore undetectable by any experiment. This is physically unacceptable because we don't consider phenomena (e.g., invisible fairies) to be of physical interest if they are undetectable even in principle.

What about the case of a material particle, i.e., one having mass? Since we already have an equation $E = m\gamma$ for the energy of a material particle in terms of its velocity, we can find a similar equation for the momentum,

$$\begin{aligned} p &= \sqrt{E^2 - m^2} \\ &= m\sqrt{\gamma^2 - 1} \\ &= m\sqrt{\frac{1}{1-v^2} - 1} \\ &= m\gamma v \end{aligned}$$

(a relation that is useful in its own right, and has been verified experimentally, f). As a material particle gets closer and closer to



f / Two early high-precision tests of the relativistic equation $p = m\gamma v$ for the momentum of a material particle. Graphing p/m rather than p allows the data for electrons and protons to be placed on the same graph. The very small error bars for the data point from Zrelov are represented by the height of the black rectangle.

c , its momentum approaches infinity, so that an infinite force would be required in order to reach c .

In summary, massless particles always move at $v = c$, while massive ones always move at $v < c$.

Note that the equation $p = m\gamma v$ isn't general enough to serve as a definition of momentum, since it becomes an indeterminate form in the limit $m \rightarrow 0$.

No half-life for massless particles *Example 11*

When we describe an unstable nucleus or other particle as having some half-life, we mean its half-life in its own rest frame. A massless particle always moves at c and therefore has no rest frame (section 3.4), so it doesn't make sense to describe it as having a half-life in this sense. This is almost, but not quite, the same thing as saying that massless particles can never decay.⁴

Constraints on polarization *Example 12*

We observe that electromagnetic waves are always polarized transversely, never longitudinally. Such a constraint can only apply to a wave that propagates at c . If it applied to a wave that propagated at less than c , we could move into a frame of reference in which the wave was at rest. In this frame, all directions in space would be equivalent, and there would be no way to decide which directions of polarization should be permitted.

4.3.5 Evidence as to which particles are massless

Which of the fundamental particles are massless, and which are not? This can only be determined empirically, and we have at least one example, the neutrino, that was formerly thought to be massless but is now believed to be massive. For more about the neutrino, see section 4.7.2, p. 109. In the present section we discuss bounds on the masses of the photon and the graviton.⁵ We omit a discussion of the gluon, which would be complicated by the fact that the gluon is never observed as a free particle or as a classical field. This section can be skipped without loss of continuity.

Some readers may exclaim at this point that of course photons must be massless, because light has to travel at the speed of light. But it should be clear from the foregoing presentation that the c in relativity is not to be interpreted as the speed of light, but as a

⁴See Fiore and Modanese, arxiv.org/abs/hep-th/9508018, and <http://physics.stackexchange.com/questions/12488/decay-of-massless-particles>. If such a process does exist, then Lorentz invariance requires that its time-scale be proportional to the particle's energy. It can be argued that gluons, which are massless, do in fact undergo decay into less energetic gluons, but the interpretation is ambiguous because we never observe gluons as free particles, so we can't just capture one in a box and watch it rattle around inside until it decays.

⁵For an in-depth review of this topic, see Goldhaber and Nieto, "Photon and Graviton Mass Limits," <http://arxiv.org/abs/0809.1003>.

kind of conversion factor between space and time. If photons have a small but nonvanishing mass, relativity does not have a stake driven through its heart.

If we want to test whether the photon is massless, the most straightforward technique would seem to be to measure its time of flight as it travels some distance, and see if it goes slower than c . There is a difficulty here because our methods for measuring large distances, e.g., GPS, generally *assume* that light travels at c . However, if the photon has some mass, then its velocity should depend on its energy, so we can instead test whether the speed of a photon depends on its energy. From quantum mechanics, this is related to its frequency by $E = hf$, so we are essentially testing whether the speed of light in a vacuum depends on frequency. Presently the best experimental tests of the invariance of the speed of light with respect to wavelength come from astronomical observations of gamma-ray bursts, which are sudden outpourings of high-energy photons, believed to originate from a supernova explosion in another galaxy. One such observation, in 2009,⁶ collected photons from such a burst, with a duration of 2 seconds, indicating that the propagation time of all the photons differed by no more than 2 seconds out of a total time in flight on the order of ten billion years, or about one part in 10^{17} !

It turns out, however, that the limits on the mass of the photon imposed by time of flight measurements can be improved on by many orders of magnitude using other methods. In the standard model of particle physics, forces are transmitted by the exchange of particles. We'll concentrate here on static forces. An electrostatic force is transmitted by the exchange of photons, and a static gravitational force by the exchange of gravitons. Gravity is not part of the standard model of particle physics, and individual gravitons cannot be directly detected by any foreseeable technology,⁷ but there are fundamental reasons for believing that they must exist, and in any case our discussion is mathematically identical for gravity and electromagnetism. We will therefore discuss electromagnetic fields and then note the corresponding results for gravity.

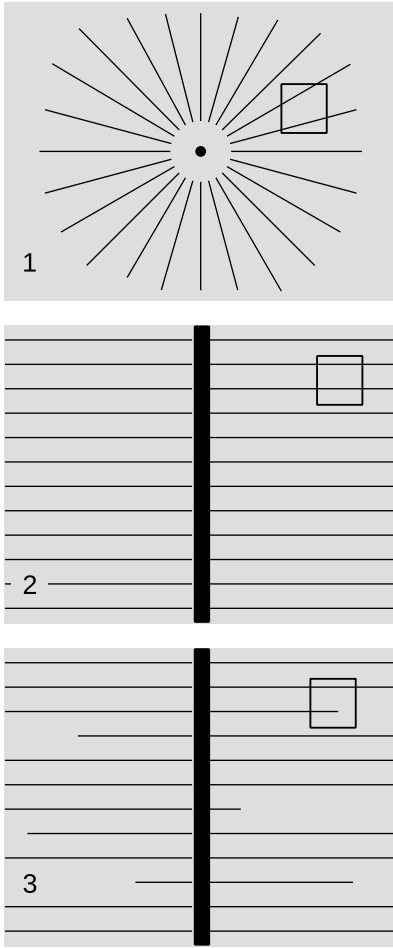
If we imagine the field surrounding a stationary point charge as a swarm of photons, then the first question that occurs to us is what is the source of the energy needed in order to create them. The standard hand-waving argument is as follows. In addition to the usual momentum-energy form of the Heisenberg uncertainty principle $\Delta p \Delta x \gtrsim h$, there is an energy-time form $\Delta E \Delta t \gtrsim h$. This looks obvious by analogy when we consider that relativistically, energy and momentum are different parts of the energy-momentum four-vector, and likewise for time and position. We can interpret this



g / An artist's conception of a gamma-ray burst, resulting from a supernova explosion.

⁶<http://arxiv.org/abs/0908.1832>

⁷Rothman and Boughn, "Can Gravitons Be Detected?," <http://arxiv.org/abs/gr-qc/0601043>



h/1. Field lines of a point charge. Observations within the small region indicated by a box allow one to determine how far away the charge is. 2. Field lines of an infinite capacitor plate, according to standard electromagnetism. Observations within the box do not give information about how far away the charge is. 3. A violation of Gauss's law.

to mean that it is possible, for short periods of time, to cheat the law of conservation of energy. We can steal a little energy but then pay it back immediately, as long as the duration of the loan is no more than about $t \sim h/E$. During this time, a virtual particle can travel a distance of no more than $\sim hc/E$. Now for a massless particle, this energy can be as small as desired, so the force can reach to arbitrarily large distances. But for a massive particle, we have the relativistic relation $E^2 - p^2 = m^2$, which requires $E \geq m$, or $E \geq mc^2$ in SI units. This minimum energy corresponds to a maximum range $\sim h/mc$. In general, we expect that the field carried by a massive particle will fall off more quickly with distance than the field of a massless particle, and we expect that this fall-off will be parametrized somehow by a length scale h/mc .

How would we expect this to play out in the classical theory of electromagnetic fields? Consider a point charge, figure h/1. Its field lines are straight, and they spread out in all directions, so by observations of any region of space, we can trace the lines backward to see where they would have intersected. That is how far it is from our region of space to the charge. This is a kind of parallax measurement. In the case of gravity, this is exactly what Eratosthenes did in order to measure the radius of the earth.

But now let's consider the case of an infinite, plane capacitor plate with some charge on it, h/2. The field lines don't spread, so the parallax method doesn't work. If we examine the field in some small region of space, there should be no way to determine the distance to the capacitor plate. If we believe in Gauss's law, then the solution is simple: the field is constant in both magnitude and direction, so although it tells us the direction of the nearest point on the plate, it tells us nothing about the distance to that point.

But if the photon is massive, we expect fields to fall off more rapidly with distance than they would according to standard theory. In this example, the standard theory says that the field does not decrease at all with distance, so for a massive photon we expect that it *does* fall off. This will violate Gauss's law, but we still expect that the distance to the plate will not be determinable by examination of a small region of space: if the field equations are linear, then a field with a given strength could be from a nearby capacitor plate with a small charge density, or a more distant one with more charge.

If we're willing to violate Gauss's law, then we can have field lines simply terminate in empty space, h/3, and this will cause the field strength to decrease. As we traverse a small distance dx , moving away from the plate, some fraction of the field lines should terminate, leading to a corresponding fractional reduction dE/E in the field strength. The ratio $(dE/E)/dx$ must be constant, and this can only happen if we have $E \propto e^{-\mu x}$, where μ is a constant with units of inverse length. (On the other side of the plate, where x is

negative, we have $+\mu x$ inside the exponential.) For the reasons discussed above, we actually expect that μ equals mc/\hbar multiplied by a unitless constant of order unity. In fact, it can be shown that the unitless constant is a factor of 2π , so μ simply the mass, expressed in units where both c and \hbar equal 1.

Since the field of a capacitor plate is equal to the superposition of the fields of all the charges distributed uniformly on it, our result that the capacitor's field falls off in a certain way tells us something corresponding about the field of a point charge. We expect that the field of a point charge q is

$$E = kq \frac{e^{-\mu r}}{r^2},$$

where Coulomb's law is recovered in the case $\mu = 0$. This form was originally inferred by Yukawa for nuclear forces, which really do have a finite range.

We now have an extraordinarily sensitive way of placing a limit on the masses of the photon and graviton. Even if μ is very small, we can make observations on very large distance scales, and static forces should fall off exponentially. In the case of gravitational forces, we observe that gravity does operate, with no detectable Yukawa-style attenuation, on scales comparable to the size of the observable universe, on the order of billions of light-years. This corresponds to a limit on the mass of the graviton of $\sim 10^{-69}$ kg — surely the smallest mass scale that has ever been probed by human beings! Measurements of the magnetic field of Jupiter by the Pioneer 10 space probe limit the mass of the photon to no more than about 8×10^{-52} kg, which is almost as impressive.

Although today's tightest bounds are from solar-system and cosmological measurements, historically some very precise tabletop experiments were carried out. Laboratory experiments are always desirable in such cases because the conditions can be controlled, and the experiments can be replicated. Problem 21 on p. 115 is an analysis of such an experiment.

4.3.6 No global conservation of energy-momentum in general relativity

If you read optional chapter 2, you know that the distinction between special and general relativity is defined by the flatness of spacetime, and that flatness is in turn defined by the path-independence of parallel transport. Whereas energy is a scalar in Newtonian mechanics, in relativity it is the timelike component of a vector. It therefore follows that in general relativity we should not expect to have global conservation of energy. For a conservation law is a statement that when we add up a certain quantity, the total has a constant value. But if spacetime is curved, then there is no natural, uniquely defined way to compare vectors that are defined at

different places in spacetime. We could parallel transport one over to the other, but the result would depend on the path along which we chose to transport it. For similar reasons, we should not expect global conservation of momentum.

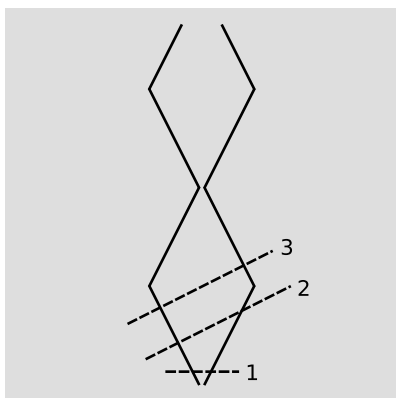
This is the answer to a frequently asked question about cosmology. Since 1998 we've known that the expansion of the universe is accelerating, rather than decelerating as we would have expected due to gravitational attraction. What is the source of the ever-increasing kinetic energy of all those galaxies? The question assumes that energy must be conserved on cosmological scales, but that just isn't so.

Nevertheless, general relativity reduces to special relativity on scales small enough to make curvature effects negligible. Therefore it is still valid to expect conservation of energy and momentum to hold *locally*, as assumed, e.g., in the analysis of Compton scattering in example 9 on p. 91, and verified in countless experiments. Cf. section 9.2, p. 179, on the stress-energy tensor.

4.4 ★ Systems with internal structure

Section 4.2 presented essentially Einstein's original proof of $E = mc^2$, which has been criticized on several grounds. A detailed discussion is given by Ohanian.⁸ Putting aside questions that are purely historical or concerned only with academic priority, we would like to know whether the proof has logical flaws, and also whether the claimed result is only valid under certain conditions. We need to consider the following questions:

1. Does it matter whether the system being described has finite spatial extent, or whether the system is isolated?
2. Does it matter whether parts of the system are moving at relativistic velocities?
3. Does the low-velocity approximation used in Einstein's proof make a difference?
4. How do we handle a system that is not made out of point-like particles, e.g., a capacitor, in which some of the energy-momentum is in an electric field?



i / The world lines of two beads bouncing back and forth on a wire.

The following example demonstrates issues 1-3 and their logical connections; the definitional question 4 is addressed in ch. 9. Suppose that two beads slide freely on a wire, bouncing elastically off of each other and also rebounding elastically from the wire's ends. Their world-lines are shown in figure i. Let's say the beads each

⁸"Einstein's $E = mc^2$ mistakes," arxiv.org/abs/0805.1400

have unit mass. In frame \mathbf{o} , the beads are released from the center of the wire with velocities $\pm u$. For concreteness, let's set $u = 1/2$, so that the system has internal motion at relativistic speeds. In this frame, the total energy-momentum vector of the system, on the surface of simultaneity labeled 1 in figure i, is $\mathbf{p} = (2.31, 0)$. That is, it has a total mass-energy of 2.31 units, and a total momentum of zero (meaning that this is the center of mass frame). As time goes on, an observer in this frame will say that the balls reach the ends of the wire simultaneously, at which point they rebound, maintaining the same total energy-momentum vector \mathbf{p} . The mass of the system is, by definition, $m = \sqrt{p_t^2 - p_x^2} = \sqrt{2.31}$, and this mass remains constant as the balls bounce back and forth.

Now let's transform into a frame \mathbf{o}' , moving at a velocity $v = 1/2$ relative to \mathbf{o} . If velocities added linearly in relativity, then the initial velocities of the beads in this frame would be 0 and -1 , but of course a material object can't move with speed $|v| = c = 1$, and velocities *don't* add linearly. Applying the correct velocity addition formula for relativity, we find that the beads have initial velocities 0 and -0.8 in this frame, and if we compute their total energy-momentum vector, on surface of simultaneity 2 in figure i, we get $\mathbf{p}' = (2.67, -1.33)$. This is exactly what we would have gotten by taking the original vector \mathbf{p} and pushing it through a Lorentz transformation. That is, the energy-momentum vector seems to be acting like a good four-vector, even though the system has finite spatial extent and contains parts that move at relativistic speeds. In particular, this implies that the system has the same mass $m = \sqrt{2.31}$ as in \mathbf{o} , since m is the norm of the \mathbf{p} vector, and the norm of a vector stays the same under a Lorentz transformation.

But now consider surface 3, which, like 2, observer \mathbf{o}' considers to be a surface of simultaneity. At this time, \mathbf{o}' says that *both* beads are moving to the left. Between time 2 and time 3, \mathbf{o}' says that the system's total momentum has changed, while its total mass-energy stayed constant. Its mass is different, and the total energy-momentum vector \mathbf{p}' at time 3 is *not* related by a Lorentz transformation to the value of \mathbf{p} at any time in frame \mathbf{o} . The reason for this misbehavior is that the right-hand bead has bounced off of the right end of the wire, but because \mathbf{o} and \mathbf{o}' have different opinions about simultaneity, \mathbf{o}' says that there has not yet been any matching collision for the bead on the left.

But all of these difficulties arise only because we have left something out. When the right-hand bead bounces off of the right-hand end of the wire, this is a collision between the bead and the wire. After the collision, the wire rebounds to the right (or a vibration is created in it). By ignoring the rebound of the wire, we have violated the law of conservation of momentum. If we take into account the momentum imparted to the wire, then the energy-momentum vector of the whole system is conserved, and must therefore be the

same at 2 and 3.

The upshot of all this is that $E = mc^2$ and the four-vector nature of \mathbf{p} are both valid for systems with finite spatial extent, provided that the systems are isolated. “Isolated” means simply that we should not gratuitously ignore anything such as the wire in this example that exchanges energy-momentum with our system. To give a general *proof* of this, it will be helpful to develop the idea of the stress-energy tensor (section 9.2, p. 179), which allows a succinct statement of what we mean by conservation of energy-momentum (subsection 9.2.1). A proof is given in section 9.3.4 on p. 191.

4.5 ★ Force

Force is a concept that is seldom needed in relativity, and that’s why this section is optional.

4.5.1 Four-force

By analogy with Newtonian mechanics, we define a relativistic force vector

$$\mathbf{F} = m\mathbf{a},$$

where \mathbf{a} is the acceleration four-vector (sec 3.5, p. 60) and m is the mass of a particle that has that acceleration as a result of the force \mathbf{F} . This is equivalent to

$$\mathbf{F} = \frac{d\mathbf{p}}{d\tau},$$

where \mathbf{p} is the mass of the particle and τ its proper time. Since the timelike part of \mathbf{p} is the particle’s mass-energy, the timelike component of the force is related to the *power* expended by the force. These definitions only work for massive particles, since for a massless particle we can’t define \mathbf{a} or τ . \mathbf{F} has been defined in terms of Lorentz invariants and four-vectors, and therefore it transforms as a god-fearing four-vector itself.

4.5.2 The force measured by an observer

The trouble with all this is that \mathbf{F} isn’t what we actually measure when we measure a force, except if we happen to be in a frame of reference that momentarily coincides with the rest frame of the particle. As with velocity and acceleration (section 3.7, p. 66), we have a four-vector that has simple, standard transformation properties, but a different \mathbf{F}_o , which is what is actually measured by the observer o . It’s defined as

$$\mathbf{F}_o = \frac{d\mathbf{p}}{dt},$$

with a dt in the denominator rather than a $d\tau$. In other words, it measures the rate of transfer of momentum according to the observer, whose time coordinate is t , not τ — unless the observer happens to be moving along with the particle. Unlike the three-vectors

\mathbf{v}_o and \mathbf{a}_o , whose timelike components are zero by definition according to observer o , \mathbf{F}_o usually has a nonvanishing timelike component, which is the rate of change of the particle's mass-energy, i.e., the power. We can refer to the spacelike part of \mathbf{F}_o as the three-force.

The following two examples show that an object moving at relativistic speeds has less inertia in the transverse direction than in the longitudinal one. A corollary is that the three-acceleration need not be parallel to the three-force.

Circular motion

Example 13

For a particle in uniform circular motion, γ is constant, and we have

$$\mathbf{F}_o = \frac{d}{dt}(m\gamma\mathbf{v}) = m\gamma \frac{d\mathbf{v}}{dt}.$$

The particle's mass-energy is constant, so the timelike component of \mathbf{F}_o does happen to be zero in this example. In terms of the three-vectors \mathbf{v}_o and \mathbf{a}_o defined in section 3.7, we have

$$\mathbf{F}_o = m\gamma \frac{d\mathbf{v}_o}{dt} = m\gamma \mathbf{a}_o,$$

which is greater than the Newtonian value by the factor γ . As a practical example, in a cathode ray tube (CRT) such as the tube in an old-fashioned oscilloscope or television, a beam of electrons is accelerated up to relativistic speed (problem 2, p. 111). To paint a picture on the screen, the beam has to be steered by transverse forces, and since the deflection angles are small, the world-line of the beam is approximately that of uniform circular motion. The force required to deflect the beam is greater by a factor of γ than would have been expected according to Newton's laws.

Linear motion

Example 14

For accelerated linear motion in the x direction, ignoring y and z , we have a velocity vector

$$\mathbf{v} = \frac{d\mathbf{r}}{d\tau},$$

whose x component is γv . Then

$$\begin{aligned} F_{o,x} &= m \frac{d(\gamma v)}{dt} \\ &= m \frac{d\gamma}{dt} v + m\gamma \frac{dv}{dt} \\ &= m \frac{d\gamma}{dv} \frac{dv}{dt} + m\gamma a \\ &= m(v^2 \gamma^3 a + \gamma a) \\ &= m\gamma^3 a \end{aligned}$$

The particle's apparent inertia is increased by a factor of γ^3 due to relativity.

The results of examples 13 and 14 can be combined as follows:

$$\mathbf{F}_{\mathbf{o}} = m\gamma \mathbf{a}_{\mathbf{o},\perp} + m\gamma^3 \mathbf{a}_{\mathbf{o},\parallel},$$

where the subscripts \perp and \parallel refer to the parts of $\mathbf{a}_{\mathbf{o}}$ perpendicular and parallel to $\mathbf{v}_{\mathbf{o}}$.

4.5.3 Transformation of the force measured by an observer

Define a frame of reference \mathbf{o} for the inertial frame of reference of an observer who does happen to be moving along with the particle at a particular instant in time. Then t is the same as τ , and $\mathbf{F}_{\mathbf{o}}$ the same as \mathbf{F} . In this frame, the particle is momentarily at rest, so the work being done on it vanishes, and the timelike components of $\mathbf{F}_{\mathbf{o}}$ and \mathbf{F} are both zero.

Suppose we do a Lorentz transformation from \mathbf{o} to a new frame \mathbf{o}' , and suppose the boost is parallel to $\mathbf{F}_{\mathbf{o}}$ and \mathbf{F} (which are both purely spatial in frame \mathbf{o}). Call this direction x . Then $d\mathbf{p} = (dp_t, dp_x) = (0, dp_x)$ transforms to $d\mathbf{p}' = (-\gamma v dp_x, \gamma dp_x)$, so that $F_{\mathbf{o}',x} = dp'_x/dt' = (\gamma dp_x)/(\gamma dt) = F_{\mathbf{o},x}$. The two factors of γ cancel, and we find that $F_{\mathbf{o}',x} = F_{\mathbf{o},x}$.

Now let's do the case where the boost is in the y direction, perpendicular to the force. The Lorentz transformation doesn't change dp_y , so $F_{\mathbf{o}',y} = dp'_y/dt' = dp_y/(\gamma dt) = F_{\mathbf{o},y}/\gamma$.

The summary of our results is as follows. Let $\mathbf{F}_{\mathbf{o}}$ be the force acting on a particle, as measured in a frame instantaneously comoving with the particle. Then in a frame of reference moving relative to this one, we have

$$\begin{aligned} F_{\mathbf{o}',\parallel} &= F_{\mathbf{o},\parallel} & \text{and} \\ F_{\mathbf{o}',\perp} &= \frac{F_{\mathbf{o},\perp}}{\gamma}, \end{aligned}$$

where \parallel indicates the direction parallel to the relative velocity of the two frames, and \perp a direction perpendicular to it.

4.5.4 Work

Consider the one-dimensional version of the three-force, $F = dp/dt$. An advantage of this quantity is that it allows us to use the Newtonian form of the (one-dimensional) work-kinetic energy relation $dE/dx = F$ without correction. Proof:

$$\begin{aligned} \frac{dE}{dx} &= \frac{dE}{dp} \frac{dp}{dt} \frac{dt}{dx} \\ &= \frac{dE}{dp} \frac{F}{v} \end{aligned}$$

By implicit differentiation of the definition of mass, we find that $dE/dp = p/E$, and this in turn equals v by the identity proved in example 7, p. 90. This leads to the claimed result, which is valid for both massless and material particles.

4.6 ★ Two applications

4.6.1 The Stefan-Boltzmann law

In 1818, Dulong and Petit analyzed experimental data to find the empirical and totally incorrect law $P \propto \exp[T/(13.5 \text{ K})]$, relating the temperature T of a body to the power it emits as electromagnetic radiation. (To see that it must be wrong, note that it doesn't vanish at absolute zero.) It was accepted until 1884, when Boltzmann corrected a systematic error in their analysis of the data, and offered a theoretical argument for the correct law, $P \propto T^4$. This law is extremely important in a variety of applications including global warming, stellar structure, cosmology, and warming your hands by the glow of a fire. Modern physics students usually come across it as a corollary in the story of the development of the quantum theory by Planck, Einstein, *et al.*, but as we will see below, it is a purely classical result, depending only on relativity and thermodynamics.

Consider an insulated cubical box of volume V containing radiation in thermal equilibrium. We let it expand uniformly with constant entropy, so that all three sides grow by the same factor a . (This is exactly what happens in cosmological expansion.) Boltzmann's clever idea was that the radiation could be treated like the working fluid in a heat engine.

By the relativistic relation between momentum and energy, the energy and momentum of a ray of light are equal (in natural units). Therefore if we lived in a one-dimensional world, the pressure p exerted by our radiation on the walls of its one-dimensional vessel would equal its energy density ρ . Because we live in a three-dimensional world, and the momenta along the three axes are in equilibrium, we have instead $p = \rho/3$. This is called the equation of state of the radiation. In cosmology, other components of the universe, such as galaxies, have equations of state with some factor other than $1/3$ in front.

As the box expands, the pressure of the radiation on the walls does work W . By conservation of energy, we have $dU + dW = 0$, where U is the energy of the radiation. Substituting $U = \rho V$ and $dW = p dV$, we obtain $d(\rho V) + p dV = 0$. Applying the product rule, separating variables, and integrating, we find $\rho \propto a^{-4}$. Here the exponent 4 is simply the number of spatial dimensions plus one. Exactly the same relation held in the early universe, which was dominated by radiation rather than matter.

Because there is no heat transfer, the entropy is constant. Entropy can be interpreted as a measure of the number of accessible states, and because state-counting doesn't depend on scaling, the occupied modes of vibration stay the same. Thus, the wavelengths simply grow in proportion to a . (A more formal and rigorous version of this argument is called the adiabatic theorem, proved by

Born and Fock in 1928.) Although this is a classical argument, we can save some work at this point by appealing to quantum mechanics for a shortcut. Since a photon has an energy $1/\lambda$, we have $U \propto 1/\lambda \propto 1/a$. The temperature of the radiation is proportional to the average energy per degree of freedom, so we have $T \propto 1/a$ as well.

Therefore $\rho \propto T^4$. This is equivalent to the Stefan-Boltzmann result, because light rays travel at the fixed speed c , and therefore the flux of radiation is proportional to the energy density. Even though this final proportionality is classical in nature, the value of the proportionality constant depends on Planck's constant, and is quantum-mechanical. A derivation is given, for example, in the Feynman Lectures on Physics, section I-41.

4.6.2 Degenerate matter

The properties of the momentum vector have surprising implications for matter subject to extreme pressure, as in a star that uses up all its fuel for nuclear fusion and collapses. These implications were initially considered too exotic to be taken seriously by astronomers.

An ordinary, smallish star such as our own sun has enough hydrogen to sustain fusion reactions for billions of years, maintaining an equilibrium between its gravity and the pressure of its gases. When the hydrogen is used up, it has to begin fusing heavier elements. This leads to a period of relatively rapid fluctuations in structure. Nuclear fusion proceeds up until the formation of elements as heavy as oxygen ($Z = 8$), but the temperatures are not high enough to overcome the strong electrical repulsion of these nuclei to create even heavier ones. Some matter is blown off, but finally nuclear reactions cease and the star collapses under the pull of its own gravity.

To understand what happens in such a collapse, we have to understand the behavior of gases under very high pressures. In general, a surface area A within a gas is subject to collisions in a time t from the n particles occupying the volume $V = Avt$, where v is the typical velocity of the particles. The resulting pressure is given by $P \sim npv/V$, where p is the typical momentum.

Nondegenerate gas: In an ordinary gas such as air, the particles are nonrelativistic, so $v = p/m$, and the thermal energy per particle is $p^2/2m \sim kT$, so the pressure is $P \sim nkT/V$.

Nonrelativistic, degenerate gas: When a fermionic gas is subject to extreme pressure, the dominant effects creating pressure are quantum-mechanical. Because of the Pauli exclusion principle, the volume available to each particle is $\sim V/n$, so its wavelength is no more than $\sim (V/n)^{1/3}$, leading to

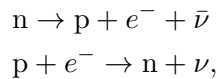
$p = h/\lambda \sim h(n/V)^{1/3}$. If the speeds of the particles are still nonrelativistic, then $v = p/m$ still holds, so the pressure becomes $P \sim (h^2/m)(n/V)^{5/3}$.

Relativistic, degenerate gas: If the compression is strong enough to cause highly relativistic motion for the particles, then $v \approx c$, and the result is $P \sim hc(n/V)^{4/3}$.

As a star with the mass of our sun collapses, it reaches a point at which the electrons begin to behave as a degenerate gas, and the collapse stops. The resulting object is called a white dwarf. A white dwarf should be an extremely compact body, about the size of the Earth. Because of its small surface area, it should emit very little light. In 1910, before the theoretical predictions had been made, Russell, Pickering, and Fleming discovered that 40 Eridani B had these characteristics. Russell recalled: “I knew enough about it, even in these paleozoic days, to realize at once that there was an extreme inconsistency between what we would then have called ‘possible’ values of the surface brightness and density. I must have shown that I was not only puzzled but crestfallen, at this exception to what looked like a very pretty rule of stellar characteristics; but Pickering smiled upon me, and said: ‘It is just these exceptions that lead to an advance in our knowledge,’ and so the white dwarfs entered the realm of study!”

S. Chandrasekhar showed in that 1930’s that there was an upper limit to the mass of a white dwarf. We will recapitulate his calculation briefly in condensed order-of-magnitude form. The pressure at the core of the star is $P \sim \rho g r \sim GM^2/r^4$, where M is the total mass of the star. The star contains roughly equal numbers of neutrons, protons, and electrons, so $M = Knm$, where m is the mass of the electron, n is the number of electrons, and $K \approx 4000$. For stars near the limit, the electrons are relativistic. Setting the pressure at the core equal to the degeneracy pressure of a relativistic gas, we find that the Chandrasekhar limit is $\sim (hc/G)^{3/2}(Km)^{-2} = 6M_\odot$. A less sloppy calculation gives something more like $1.4M_\odot$.

What happens to a star whose mass is above the Chandrasekhar limit? As nuclear fusion reactions flicker out, the core of the star becomes a white dwarf, but once fusion ceases completely this cannot be an equilibrium state. Now consider the nuclear reactions



which happen due to the weak nuclear force. The first of these releases 0.8 MeV, and has a half-life of 14 minutes. This explains why free neutrons are not observed in significant numbers in our universe, e.g., in cosmic rays. The second reaction requires an *input* of 0.8 MeV of energy, so a free hydrogen atom is stable. The white



j / Subrahmanyan Chandrasekhar (1910-1995)

dwarf contains fairly heavy nuclei, not individual protons, but similar considerations would seem to apply. A nucleus can absorb an electron and convert a proton into a neutron, and in this context the process is called electron capture. Ordinarily this process will only occur if the nucleus is neutron-deficient; once it reaches a neutron-to-proton ratio that optimizes its binding energy, neutron capture cannot proceed without a source of energy to make the reaction go. In the environment of a white dwarf, however, there is such a source. The annihilation of an electron opens up a hole in the “Fermi sea.” There is now an state into which another electron is allowed to drop without violating the exclusion principle, and the effect cascades upward. In a star with a mass above the Chandrasekhar limit, this process runs to completion, with every proton being converted into a neutron. The result is a *neutron star*, which is essentially an atomic nucleus (with $Z = 0$) with the mass of a star!

Observational evidence for the existence of neutron stars came in 1967 with the detection by Bell and Hewish at Cambridge of a mysterious radio signal with a period of 1.3373011 seconds. The signal’s observability was synchronized with the rotation of the earth relative to the stars, rather than with legal clock time or the earth’s rotation relative to the sun. This led to the conclusion that its origin was in space rather than on earth, and Bell and Hewish originally dubbed it LGM-1 for “little green men.” The discovery of a second signal, from a different direction in the sky, convinced them that it was not actually an artificial signal being generated by aliens. Bell published the observation as an appendix to her PhD thesis, and it was soon interpreted as a signal from a neutron star. Neutron stars can be highly magnetized, and because of this magnetization they may emit a directional beam of electromagnetic radiation that sweeps across the sky once per rotational period — the “lighthouse effect.” If the earth lies in the plane of the beam, a periodic signal can be detected, and the star is referred to as a pulsar. It is fairly easy to see that the short period of rotation makes it difficult to explain a pulsar as any kind of less exotic rotating object. In the approximation of Newtonian mechanics, a spherical body of density ρ , rotating with a period $T = \sqrt{3\pi/G\rho}$, has zero apparent gravity at its equator, since gravity is just strong enough to accelerate an object so that it follows a circular trajectory above a fixed point on the surface (problem 17). In reality, astronomical bodies of planetary size and greater are held together by their own gravity, so we have $T \gtrsim 1/\sqrt{G\rho}$ for any body that does not fly apart spontaneously due to its own rotation. In the case of the Bell-Hewish pulsar, this implies $\rho \gtrsim 10^{10} \text{ kg/m}^3$, which is far larger than the density of normal matter, and also 10-100 times greater than the typical density of a white dwarf near the Chandrasekhar limit.

An upper limit on the mass of a neutron star can be found in a manner entirely analogous to the calculation of the Chandrasekhar

limit. The only difference is that the mass of a neutron is much greater than the mass of an electron, and the neutrons are the only particles present, so there is no factor of K . Assuming the more precise result of $1.4M_\odot$ for the Chandrasekhar limit rather than our sloppy one, and ignoring the interaction of the neutrons via the strong nuclear force, we can infer an upper limit on the mass of a neutron star:

$$1.4M_\odot \left(\frac{Km_e}{m_n} \right)^2 \approx 5M_\odot$$

The theoretical uncertainties in such an estimate are fairly large. Tolman, Oppenheimer, and Volkoff originally estimated it in 1939 as $0.7M_\odot$, whereas modern estimates are more in the range of 1.5 to $3M_\odot$. These are significantly lower than our crude estimate of $5M_\odot$, mainly because the attractive nature of the strong nuclear force tends to pull the star toward collapse. Unambiguous results are presently impossible because of uncertainties in extrapolating the behavior of the strong force from the regime of ordinary nuclei, where it has been relatively well parametrized, into the exotic environment of a neutron star, where the density is significantly different and no protons are present. There are a variety of effects that may be difficult to anticipate or to calculate. For example, Brown and Bethe found in 1994⁹ that it might be possible for the mass limit to be drastically revised because of the process $e^- \rightarrow K^- + \nu_e$, which is impossible in free space due to conservation of energy, but might be possible in a neutron star. Observationally, nearly all neutron stars seem to lie in a surprisingly small range of mass, between 1.3 and $1.45M_\odot$, but in 2010 a neutron star with a mass of $1.97 \pm .04 M_\odot$ was discovered, ruling out most neutron-star models that included exotic matter.¹⁰

For stars with masses above the Tolman-Oppenheimer-Volkoff limit, it seems likely, both on theoretical and observational grounds, we end up with a black hole: an object with an event horizon (cf. p. 62) that cuts its interior off from the rest of the universe.

4.7 ★ Tachyons and FTL

4.7.1 A defense in depth

Let's summarize some ideas about faster-than-light (FTL, superluminal) motion in relativity:

1. Superluminal transmission of information would violate causality, since it would allow a causal relationship between events

⁹H.A. Bethe and G.E. Brown, "Observational constraints on the maximum neutron star mass," *Astrophys. J.* 445 (1995) L129. G.E. Brown and H.A. Bethe, "A Scenario for a Large Number of Low-Mass Black Holes in the Galaxy," *Astrophys. J.* 423 (1994) 659. Both papers are available at adsabs.harvard.edu.

¹⁰Demorest et al., arxiv.org/abs/1010.5788v1.

that were spacelike in relation to one another, and the time-ordering of such events is different according to different observers. Since we never seem to observe causality to be violated, we suspect that superluminal transmission of information is impossible. This leads us to interpret the metric in relativity as being fundamentally a statement of possible cause and effect relationships between events.

2. We observe the invariant mass defined by $m^2 = E^2 - p^2$ to be a fixed property of all objects. Therefore we suspect that it is not possible for an object to change from having $|E| > |p|$ to having $|E| < |p|$.
3. No continuous process of acceleration can bring an observer from $v < c$ to $v > c$ (see section 3.3). Since it's possible to build an observer out of material objects, it seems that it's impossible to get a material object past c by a continuous process of acceleration.
4. If superluminal motion were possible, then one might also expect superluminal observers to be possible. But FTL frames of reference are kinematically impossible in $3 + 1$ dimensions (section 3.8, p. 69).

Thus special relativity seems to have a defense in depth against superluminal motion.

Based on 2, FTL motion would be a property of an exotic form of matter built out of hypothetical particles with imaginary mass. Such particles are called tachyons. An imaginary mass is not absurd on its face, because experiments directly measure E and p , not m . E.g., if we put a tachyon on a scale and weighed it, we would be measuring its mass-energy E .

The weakest of these arguments is 1, since as described in section 2.1, we have no strong reasons for believing in causality as an overarching principle of physics. It would be exciting if we could detect tachyons in particle accelerator experiments or as naturally occurring radiation. Perhaps we could even learn to transmit and receive tachyon signals artificially, allowing us to send ourselves messages from the future! This possibility was pointed out in 1917 by Tolman¹¹ and is referred to as the “tachyonic antitelephone.”¹²

¹¹www.archive.org/details/theoryrelativmot00tolmrch

¹²Bilaniuk *et al.* claimed in a 1962 paper to have found a reinterpretation that eliminated the causality violation, but their interpretation requires that rates of tachyon emission in one frame be related to rates of tachyon absorption in another frame, which in my opinion is equally problematic, since rates of absorption should depend on the environment, whereas rates of emission should depend on the emitter; the causality violation has simply been described in different words, but not eliminated. For a different critique, see Benford, Book, and Newcomb, “The tachyonic antitelephone,” *Physical Review D* 2 (1970) 263. Scans of the paper can be found online.

If we're willing to let go of causality, then we just need to make sure that our tachyons comply with items 3 and 4 above. Argument 4 tells us that the laws of physics must conspire to make it impossible to build an observer out of tachyons; this is not entirely implausible, since there are other classes of particles such as photons that can't be used to construct observers.

4.7.2 Experiments to search for tachyons

Experimental searches are made more difficult by conflicting theoretical claims as to whether tachyons should be charged or neutral, whether they should have integral or half-integral spin, and whether the normal spin-statistics relation even applies to them.¹³ If charged, it is uncertain whether and under what circumstances they would emit Cerenkov radiation.

The most obvious experimental signature of tachyons would be propagation at speeds greater than c . Negative results were reported by Murthy and later by Clay,¹⁴ who studied air showers generated by cosmic rays to look for precursor particles that arrived before the first photons.

One could also look for particles with $|p| > |E|$. Alvager and Erman, in a 1965 experiment, studied the beta decay of ^{170}Tm , using a spectrometer to measure the momentum of charged radiation and a solid state detector to determine energy. An upper limit of one tachyon per 10^4 beta particles was inferred.

If tachyons are neutral, then they might be difficult to detect directly, but it might be possible to infer their existence indirectly through missing energy-momentum in reactions. This is how the neutrino was first discovered. Baltay *et al.*¹⁵ searched for reactions such as $\bar{p} + p \rightarrow \pi^+ + \pi^- + t$, with t being a neutral tachyon, by measuring the momenta of all the other initial and final particles and looking for events in which the missing energy-momentum was spacelike. They put upper limits of $\sim 10^{-3}$ on the branching ratios of this and several other reactions leading to production of single tachyons or tachyon-antitachyon pairs.

For a long time after the discovery of the neutrino, very little was known about its mass, so it was consistent with the experimental evidence to imagine that one or more species of neutrinos were tachyons, and Chodos *et al.* made such speculations in 1985. A brief flurry of reawakened interest in tachyons was occasioned by a 2011 debacle in which the particle-physics experiment OPERA mistakenly reported faster-than-light propagation of neutrinos; the

¹³Feinberg, "Possibility of Faster-Than-light Particles," Phys Rev 159 (1967) 1089, <http://www.scribd.com/doc/144943457/G-Feinberg-Possibility-of-Faster-Than-light-Particles-Phys-Rev-159-1967-1089>

¹⁴"A search for tachyons in cosmic ray showers," Austr. J. Phys 41 (1988) 93, <http://adsabs.harvard.edu/full/1988AuJPh...41...93C>

¹⁵Phys. Rev. D 1 (1970) 759

anomaly was later found to be the result of a loose connection on a fiber-optic cable plus a miscalibrated oscillator. An experiment called KATRIN, currently nearing the start of operation at Karlsruhe, will provide the first direct measurement of the mass of the neutrino, by measuring very precisely the maximum energy of the electrons emitted in the decay of tritium, ${}^3\text{H} \rightarrow {}^3\text{He} + e^- + \bar{\nu}_e$. Conservation of energy then allows one to determine the *minimum* energy of the antineutrino, which is related to its mass and momentum by $m^2 = E^2 - p^2$. Because m^2 appears in this equation, the experiment really measures m^2 , not m , and a result of $m^2 < 0$ would bring the tachyonic neutrino back from the grave.

4.7.3 Tachyons and quantum mechanics

When we add quantum mechanics to special relativity, we get quantum field theory, which sounds scary and can be quite technical, but is governed by some very simple principles. One of these principles is that “everything not forbidden is compulsory.” The phrase was popularized as a political satire of communism by T.H. White, but was commandeered by physicist Murray Gell-Mann to express the idea that any process not forbidden by a conservation law will in fact occur in nature at some rate. If tachyons exist, then it is possible to have two tachyons whose energy-momentum vectors add up to zero (problem 8, p. 112). This would seem to imply that the vacuum could spontaneously create tachyon-antitachyon pairs. Most theorists now interpret this as meaning that when tachyons pop up in the equations, it’s a sign that the assumed vacuum state is not stable, and will change into some other state that is the true state of minimum energy.

Problems

1 Criticize the following reasoning. *Temperature is a measure of the energy per atom. In nonrelativistic physics, there is a minimum temperature, which corresponds to zero energy per atom, but no maximum. In relativity, there should be a maximum temperature, which would be the temperature at which all the atoms are moving at c .*

2 In an old-fashioned cathode ray tube (CRT) television, electrons are accelerated through a voltage difference that is typically about 20 kV. At what fraction of the speed of light are the electrons moving?

3 In nuclear beta decay, an electron or antielectron is typically emitted with an energy on the order of 1 MeV. In alpha decay, the alpha particle typically has an energy of about 5 MeV. In each case, do a rough estimate of whether the particle is nonrelativistic, relativistic, or ultrarelativistic.

4 Suppose that the starship Enterprise from Star Trek has a mass of 8.0×10^7 kg, about the same as the Queen Elizabeth 2. Compute the kinetic energy it would have to have if it was moving at half the speed of light. Compare with the total energy content of the world's nuclear arsenals, which is about 10^{21} J. ✓

5 Cosmic-ray neutrinos may be the fastest material particles in the universe. In 2013 the IceCube neutrino detector in Antarctica detected two neutrinos,¹⁶ dubbed Bert and Ernie, after the Sesame Street characters, with energies in the neighborhood of 1 PeV = 10^{15} eV. The higher energy was Ernie's 1.14 ± 0.17 PeV. It is not known what type of neutrino he was, nor do we have exact masses for neutrinos, but let's assume $m = 1$ eV. Find Ernie's rapidity.

6 Science fiction stories often depict spaceships traveling through solar systems at relativistic speeds. Interplanetary space contains a significant number of tiny dust particles, and such a ship would sweep these dust particles out of a large volume of space, impacting them at high speeds. A 1975 experiment aboard the Skylab space station measured the frequency of impacts from such objects and found that a square meter of exposed surface experienced an impact from a particle with a mass of $\sim 10^{-15}$ kg about every few hours. A relativistic object, sweeping through space much more rapidly, would experience such impacts at rates of more like one every few seconds. (Larger particles are significantly more rare, with the frequency falling off as something like m^{-8} .) These particles didn't damage Skylab, because at relative velocities of $\sim 10^4$ m/s their kinetic energies were on the order of microjoules. At relativistic speeds it would be a different story. Real-world spacecraft are lightweight and rather fragile, so there would probably be serious consequences

¹⁶arxiv.org/abs/1304.5356

from any impact having a kinetic energy of about 10^2 J (comparable to a bullet from a small handgun). (a) Find the speed at which a starship could cruise through a solar system if frequent 10^2 J collisions were acceptable, assuming no object with a mass of more than 10^{-15} kg. Express your result relative to c . (b) Find the speed under the more conservative parameters of 10 J and 10^{-14} kg.

7 Example 4 on p. 61 derives the equation

$$x = \frac{1}{a} \cosh a\tau$$

for a particle moving with constant acceleration. (Note that a constant of integration was taken to be zero, so that $x \neq 0$ at $\tau = 0$.)

(a) Rewrite this equation in metric units by inserting the necessary factors of c . (b) If we had a rocket ship capable of accelerating indefinitely at g , how much proper time would be needed in order to travel the distance $\Delta x = 27,000$ light-years to the galactic center? (This will be a flyby, so the ship accelerates all the way rather than decelerating to stop at its destination.) Answer: 11 years (c) An observer at rest relative to the galaxy explains the surprisingly short time calculated in part b as being due to the time dilation experienced by the traveler. How does the traveler explain it?

8 Show, as claimed on p. 110, that if tachyons exist, then it is possible to have two tachyons whose momentum vectors add up to zero.

9 (a) A free neutron (as opposed to a neutron bound into an atomic nucleus) is unstable, and undergoes spontaneous radioactive decay into a proton, an electron, and an antineutrino. The masses of the particles involved are as follows:

neutron	1.67495×10^{-27} kg
proton	1.67265×10^{-27} kg
electron	0.00091×10^{-27} kg
antineutrino	$< 10^{-35}$ kg

Find the energy released in the decay of a free neutron. ✓

(b) Neutrons and protons make up essentially all of the mass of the ordinary matter around us. We observe that the universe around us has no free neutrons, but lots of free protons (the nuclei of hydrogen, which is the element that 90% of the universe is made of). We find neutrons only inside nuclei along with other neutrons and protons, not on their own.

If there are processes that can convert neutrons into protons, we might imagine that there could also be proton-to-neutron conversions, and indeed such a process does occur sometimes in nuclei that contain both neutrons and protons: a proton can decay into a neutron, a positron, and a neutrino. A positron is a particle with the same properties as an electron, except that its electrical charge is positive. A neutrino, like an antineutrino, has negligible mass.

Although such a process can occur within a nucleus, explain why it cannot happen to a free proton. (If it could, hydrogen would be radioactive, and you wouldn't exist!)

10 (a) Find a relativistic equation for the velocity of an object in terms of its mass and momentum (eliminating γ). \checkmark

(b) Show that your result is approximately the same as the classical value, p/m , at low velocities.

(c) Show that very large momenta result in speeds close to the speed of light.

11 Expand the equation for relativistic kinetic energy $K = m(\gamma - 1)$ in a Taylor series, and find the first two nonvanishing terms. Show that the first term is the nonrelativistic expression.

12 Expand the equation $p = m\gamma v$ in a Taylor series, and find the first two nonvanishing terms. Show that the first term is the classical expression.

13 An atom in an excited state emits a photon, ending up in a lower state. The initial state has mass m_1 , the final one m_2 . To a very good approximation, we expect the energy E of the photon to equal $m_1 - m_2$. However, conservation of momentum dictates that the atom must recoil from the emission, and therefore it carries away a small amount of kinetic energy that is not available to the photon. Find the exact energy of the photon, in the frame in which the atom was initially at rest.

14 The following are the three most common ways in which gamma rays interact with matter:

Photoelectric effect: The gamma ray hits an electron, is annihilated, and gives all of its energy to the electron.

Compton scattering: The gamma ray bounces off of an electron, exiting in some direction with some amount of energy.

Pair production: The gamma ray is annihilated, creating an electron and a positron.

Example 10 on p. 92 shows that pair production can't occur in a vacuum due to conservation of the energy-momentum four-vector. What about the other two processes? Can the photoelectric effect occur without the presence of some third particle such as an atomic nucleus? Can Compton scattering happen without a third particle?

15 This problem assumes you know some basic quantum physics. The point of this problem is to estimate whether or not a neutron or proton in an atomic nucleus is highly relativistic. Nuclei typically have diameters of a few fm ($1 \text{ fm} = 10^{-15} \text{ m}$). Take a neutron or proton to be a particle in a box of this size. In the ground state, half a wavelength would fit in the box. Use the de Broglie relation to estimate its typical momentum and thus its typical speed. How relativistic is it?

16 Show, as claimed in example 11 on p. 94, that if a massless particle were to decay, Lorentz invariance requires that the time-scale τ for the process be proportional to the particle's energy. What units would the constant of proportionality have?

17 Derive the equation $T = \sqrt{3\pi/G\rho}$ given on page 106 for the period of a rotating, spherical object that results in zero apparent gravity at its surface.

18 Neutrinos with energies of $\sim 1 \text{ MeV}$ (the typical energy scale of nuclear physics) make up a significant part of the matter in our universe. If a neutrino and an antineutrino annihilate each other, the product is two back-to-back photons whose energies are equal in the center-of-mass frame. Should astronomers be able to detect these photons by selecting only those with the correct energy?

19 In a certain frame of reference, a gamma ray with energy E_1 is moving to the right, while a second gamma ray with energy E_2 flies off to the left. (a) Find the mass of the system. (b) Find the velocity of the center-of-mass frame, i.e., the frame of reference in which the total momentum is zero.

20 In section 4.5.4 we proved the work-energy relation $dE/dx = F$ in the context of relativity. Recapitulate the derivation in the context of pure Newtonian mechanics.

21 Section 4.3.5 on p. 94 discusses the possibility that the photon has a small but nonzero mass m . One of the consequences is that the electric field of an infinite, uniformly charged plane is $E_x = \pm 2\pi k\sigma \exp(-\mu|x|)$, where k is the Coulomb constant, σ is the charge per unit area, $\mu = mc/\hbar$, and x is the distance from the plane. When $m = 0$, we recover the result of standard electromagnetism. The purpose of this problem is to analyze a laboratory experiment that can put an upper bound on m .

Consider a rectangular, hollow, conducting box with charge placed on it. If $m = 0$, then Gauss's law holds, and the field inside is exactly zero. We now consider the possibility that $m > 0$. We make the box very thin in the x direction, with sides located at $x = \pm a$. We refer to these two sides as the "plates." The box's extent in the y and z directions is much greater than a , so that the density of charge σ on each of the two plates is nearly constant as long as we stay away from the fringing fields at the edges. Consider a point located at $x = b$, with $0 < b < a$, and far from the edges. Show that there is a nonvanishing interior field, which can be measured in this experiment by the fractional difference in electric potential

$$\frac{V(a) - V(b)}{V(a)} \approx \frac{1}{2}m^2(a^2 - b^2) + \dots,$$

where \dots indicates higher-order terms.

Remark: The experiment is more practical when carried out using a spherical geometry, since there are no fringing fields to worry about. The analysis comes out the same except that the factor of $1/2$ becomes $1/6$. Experiments of this type were first carried out by Cavendish in 1722, and then with a series of order-of-magnitude improvements in precision by Plimpton in 1936 and Williams in 1971.

22 Potassium 40 is the strongest source of naturally occurring beta radioactivity in our environment. It decays according to



The energy released in the decay is 1.33 MeV. The energy is shared randomly among the products, subject to the constraint imposed by conservation of energy-momentum, which dictates that very little of the energy is carried by the recoiling calcium nucleus. Determine the maximum energy of the calcium, and compare with the typical energy of a chemical bond, which is a few eV. If the potassium is part of a molecule, do we expect the molecule to survive? Carry out the calculation first by assuming that the electron is ultrarelativistic, then without the approximation, and comment on the how good the approximation is.

23 The products of a certain radioactive decay are a massive particle and a gamma ray.

(a) Show that, in the center of mass frame, the energy of the gamma is less than the mass-energy of the massive particle.

(b) Show that the opposite inequality holds if we compare the *kinetic* energy of the massive particle to the energy of the gamma.

(c) Suppose someone tells you that a certain massive particle has a mode of radioactive decay in which it disappears, and the only product is a gamma ray — no residual massive particle exists. Use the result of part a to show that this is impossible, and then see if you can find a simpler argument to demonstrate the same thing. [Based on a problem by B. Shotwell.]

Chapter 5

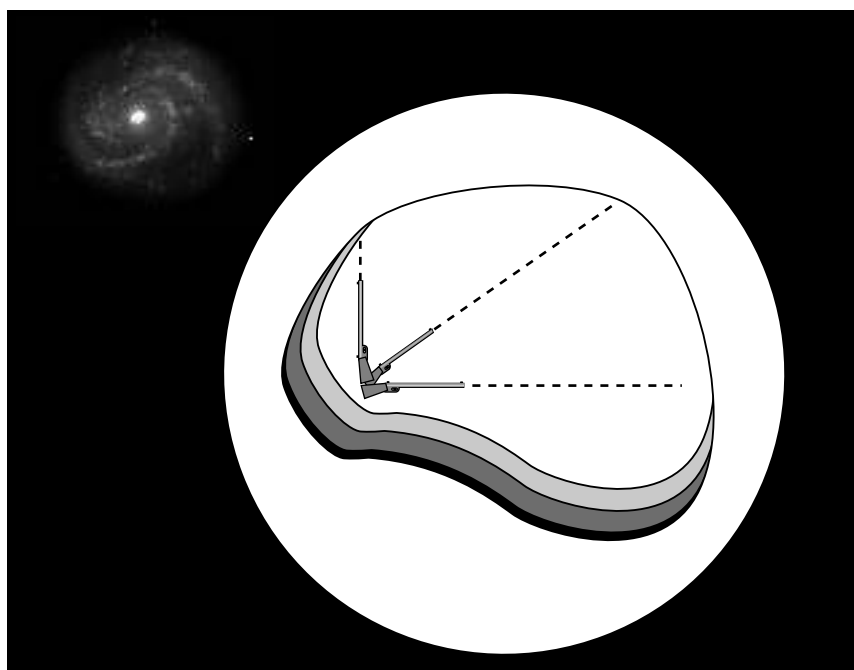
Inertia (optional)

5.1 What is inertial motion?

On p. 47 I stated the following as an axiom of special relativity:

P4. *Inertial frames of reference exist.* These are frames in which particles move at constant velocity if not subject to any forces. We can construct such a frame by using a particular particle, which is not subject to any forces, as a reference point. Inertial motion is modeled by vectors and parallelism.

This is a typical modern restatement of Newton's first law. It claims to define inertial frames and claims that they exist.



a / The spherical chamber, shown in a cutaway view, has layers of shielding to exclude all known nongravitational forces. The three guns, at right angles to each other, fire bullets. Once the chamber has been calibrated by marking the three dashed-line trajectories under free-fall conditions, an observer inside the chamber can always tell whether she is in an inertial frame.

5.1.1 An operational definition

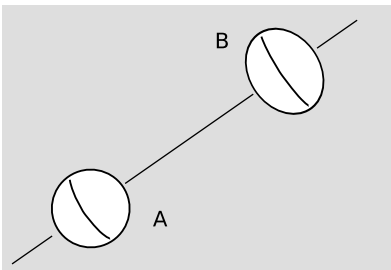
In keeping with the philosophy of operationalism (p. 26), we ought to be able to translate the definition into a method for testing whether a given frame really is inertial. Figure a shows an idealized variation on a device actually built for this purpose by Harold Waage at Princeton as a lecture demonstration to be used by his partner in

crime John Wheeler. We build a sealed chamber whose contents are isolated as much as possible from outside forces. Of the four known forces of nature, the ones we know how to exclude are the strong nuclear force, the weak nuclear force, and the electromagnetic force. The strong nuclear force has a range of only about 1 fm (10^{-15} m), so to exclude it we merely need to make the chamber thicker than that, and also surround it with enough paraffin wax to keep out any neutrons that happen to be flying by. The weak nuclear force also has a short range, and although shielding against neutrinos is a practical impossibility, their influence on the apparatus inside will be negligible. To shield against electromagnetic forces, we surround the chamber with a Faraday cage and a solid sheet of mu-metal. Finally, we make sure that the chamber is not being touched by any surrounding matter, so that short-range residual electrical forces (sticky forces, chemical bonds, etc.) are excluded. That is, the chamber cannot be supported; it is free-falling.

Crucially, the shielding does *not* exclude gravitational forces. There is in fact no known way of shielding against gravitational effects such as the attraction of other masses or the propagation of gravitational waves. (Because the shielding is spherical, it exerts no gravitational force of its own on the apparatus inside.)

Inside, an observer carries out an initial calibration by firing bullets along three Cartesian axes and tracing their paths, which she *defines* to be linear. (She can also make sure that the chamber isn't rotating, e.g., by checking for velocity-dependent Coriolis forces.) After the initial calibration, she can always tell whether or not she is in an inertial frame. She simply has to fire the bullets, and see whether or not they follow the precalibrated paths. For example, she can detect that the frame has become noninertial if the chamber is rotated, allowed to rest on the ground, or accelerated by a rocket engine.

Isaac Newton would have been extremely unhappy with our definition. "This is absurd," he says. "The way you've defined it, my street in London isn't inertial." Newtonian mechanics only makes predictions if we input the correct data on all the mass in the universe. Given this kind of knowledge, we can properly account for all the gravitational forces, and define the street in London as an inertial frame because in that frame, the trees and houses have zero total force on them and don't accelerate. But spacetime isn't Galilean. In special relativity's description of spacetime, information propagates at a maximum speed of c , so there will always be distant parts of the universe that we can never know about, because information from those regions hasn't had time to reach us yet.



b / Example 1.

Rotation is noninertial

Example 1

Figure b shows a hypothetical example proposed by Einstein. One planet rotates about its axis and therefore has an equatorial

bulge. The other planet doesn't rotate and has none. Both Newtonian mechanics and special relativity make these predictions, and although the scenario is idealized and unrealistic, there is no doubt that their predictions are correct for this situation, because the two theories have been tested in similar cases. This also agrees with our operational definition of inertial motion on p. 118. Rotational motion is noninertial.

This bothered Einstein for the following reason. If the inhabitants of the two planets can look up in the sky at the "fixed stars," they have a clear explanation of the reason for the difference in shape. People on planet A don't see the stars rise or set, and they infer that this is because they live on a nonrotating world. The inhabitants of planet B do see the stars rise and set, just as they do here on earth, so they infer, just as Copernicus did, that their planet rotates.

But suppose, Einstein said, that the two planets exist alone in an otherwise empty universe. There are no stars. Then it's equally valid for someone on either planet to say that it's the one that doesn't rotate. Each planet rotates *relative to the other planet*, but the situation now appears completely symmetric. Einstein took this argument seriously and felt that it showed a defect in special relativity. He hoped that his theory of general relativity would fix this problem, and predict that in an otherwise empty universe, neither planet would show any tidal bulge. In reality, further study of the general theory of relativity showed that it made the same prediction as special relativity. Theorists have constructed other theories of gravity, most prominently the Brans-Dicke theory, that do behave more in the way Einstein's physical intuition expected. Precise solar-system tests have, however, supported general relativity rather than Brans-Dicke gravity, so it appears well settled now that rotational motion really shouldn't be considered inertial.

5.1.2 Equivalence of inertial and gravitational mass

All of the reasoning above depends on the perfect cancellation referred to by Newton: since gravitational forces are proportional to mass, and acceleration is inversely proportional to mass, the result is that accelerations caused by gravity are independent of mass. This is the universality of free fall, which was famously observed by Galileo, figure c.

Suppose that, on the contrary, we had access to some matter that was immune to gravity. It's sold under the brand name FloatyStuff™. The cancellation fails now. Let's say that alien gangsters land in a flying saucer, kidnap you out of your back yard, konk you on the head, and take you away. When you regain consciousness, you're locked up in a sealed cabin in their spaceship. You pull your keychain out of your pocket and release it, and you



c / According to Galileo's student Viviani, Galileo dropped a cannonball and a musketball simultaneously from the leaning tower of Pisa, and observed that they hit the ground at nearly the same time. This contradicted Aristotle's long-accepted idea that heavier objects fell faster.

observe that it accelerates toward the floor with an acceleration that seems quite a bit slower than what you're used to on earth, perhaps a third of a gee. There are two possible explanations for this. One is that the aliens have taken you to some other planet, maybe Mars, where the strength of gravity is a third of what we have on earth. The other is that your keychain didn't really accelerate at all: you're still inside the flying saucer, which is accelerating at a third of a gee, so that it was really the deck that accelerated up and hit the keys.

There is absolutely no way to tell which of these two scenarios is actually the case — unless you happen to have a chunk of FloatyStuff in your other pocket. If you release the FloatyStuff and it hovers above the deck, then you're on another planet and experiencing genuine gravity; your keychain responded to the gravity, but the FloatyStuff didn't. But if you release the FloatyStuff and see it hit the deck, then the flying saucer is accelerating through outer space.

5.2 The equivalence principle

5.2.1 Equivalence of acceleration to a gravitational field

The nonexistence of FloatyStuff in our universe is a special case of the *equivalence principle*. The equivalence principle states that an acceleration (such as the acceleration of the flying saucer) is always equivalent to a gravitational field, and no observation can ever tell the difference without reference to something external. (And suppose you did have some external reference point — how would you know whether *it* was accelerating?)

5.2.2 Eötvös experiments

FloatyStuff would be an extreme example, but if there was any violation of the universality of free fall, no matter how small, then the equivalence principle would be falsified. Since Galileo's time, experimental methods have had several centuries in which to improve, and the second law has been subjected to similar tests with exponentially improving precision. For such an experiment in 1993,¹ physicists at the University of Pisa (!) built a metal disk out of copper and tungsten semicircles joined together at their flat edges. They evacuated the air from a vertical shaft and dropped the disk down it 142 times, using lasers to measure any tiny rotation that would result if the accelerations of the copper and tungsten were very slightly different. The results were statistically consistent with zero rotation, and put an upper limit of 1×10^{-9} on the fractional difference in acceleration $|g_{\text{copper}} - g_{\text{tungsten}}|/g$. Experiments of this type are called Eötvös experiments, after Loránd Eötvös, who did the first modern, high-precision versions.



d / Loránd Eötvös (1848-1919).

¹Carusotto *et al.*, "Limits on the violation of g -universality with a Galileo-type experiment," Phys Lett A183 (1993) 355. Freely available online at researchgate.net.

The artificial horizon

Example 2

The pilot of an airplane cannot always easily tell which way is up. The horizon may not be level simply because the ground has an actual slope, and in any case the horizon may not be visible if the weather is foggy. One might imagine that the problem could be solved simply by hanging a pendulum and observing which way it pointed, but by the equivalence principle the pendulum cannot tell the difference between a gravitational field and an acceleration of the aircraft relative to the ground — nor can any other accelerometer, such as the pilot's inner ear. For example, when the plane is turning to the right, accelerometers will be tricked into believing that “down” is down and to the left. To get around this problem, airplanes use a device called an artificial horizon, which is essentially a gyroscope. The gyroscope has to be initialized when the plane is known to be oriented in a horizontal plane. No gyroscope is perfect, so over time it will drift. For this reason the instrument also contains an accelerometer, and the gyroscope is automatically restored to agreement with the accelerometer, with a time-constant of several minutes. If the plane is flown in circles for several minutes, the artificial horizon will be fooled into indicating that the wrong direction is vertical.



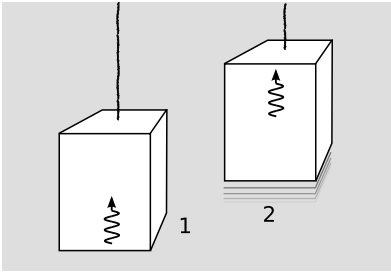
e / An artificial horizon.

5.2.3 Gravity without gravity

We live immersed in the earth's gravitational field, and that is where we do almost all of our physics experiments. It's surprising, then, that special relativity can be confirmed in earth-bound experiments, sometimes with phenomenal precision, as in the Ives-Stilwell experiment's 10-significant-figure test of the relativistic Doppler shift equation (p. 56). How can this be, since special relativity is supposed to be the version of relativity that can't handle gravity? The equivalence principle provides an answer. If the only gravitational effect on your experiment is a uniform field \mathbf{g} , then it's valid for you to describe your experiment as having been done in a region *without* any gravity, but in a laboratory whose floor happened to have been accelerating upward with an acceleration $-\mathbf{g}$. Special relativity works just fine in such situations, because switching into an accelerated frame of reference doesn't have any effect on the flatness of spacetime (p. 46). Note that Gravity Probe B (p. 46) orbited the earth, so the field it experienced varied in direction, causing the above argument to fail; the effects it observed were not explainable by special relativity.

5.2.4 Gravitational Doppler shifts

For an example of a specifically gravitational experiment that is explainable by special relativity, and that has actually been carried out, In a laboratory accelerating upward, a light wave emitted from the floor would be Doppler-shifted toward lower frequencies when observed at the ceiling, because of the change in the receiver's



1. A light wave is emitted upward from the floor of the elevator. The elevator accelerates upward. 2. By the time the light wave is detected at the ceiling, the elevator has changed its velocity, so the wave is detected with a Doppler shift.



g / Pound and Rebka at the top and bottom of the tower.

velocity during the wave's time of flight. The effect is given by $\Delta f/f \approx -a\Delta x/c^2$, where a is the lab's acceleration, Δx is the height from floor to ceiling, and c is the speed of light (problem 1). In units with $c = 1$, we have $\Delta f/f \approx -a\Delta x$.

By the equivalence principle, we find that when such an experiment is done in a gravitational field g , there should be a gravitational effect on frequency $\Delta f/f \approx -g\Delta x$. This can be expressed more compactly as $\Delta f/f \approx -\Delta\Phi$, where Φ is the gravitational potential, i.e., the gravitational energy per unit mass.

In 1959, Pound and Rebka² carried out an experiment in a tower at Harvard. Gamma rays from were emitted by a ^{57}Fe source at the bottom and detected at the top, having risen $\Delta x = 22.6$ m. The equivalence principle predicts a fractional frequency shift due to gravity of 2.46×10^{-15} . This is very small, and would normally have been masked by recoil effects (problem 13, p. 113), but by exploiting the Mössbauer effect Pound and Rebka measured the shift to be $(2.56 \pm 0.25) \times 10^{-15}$.

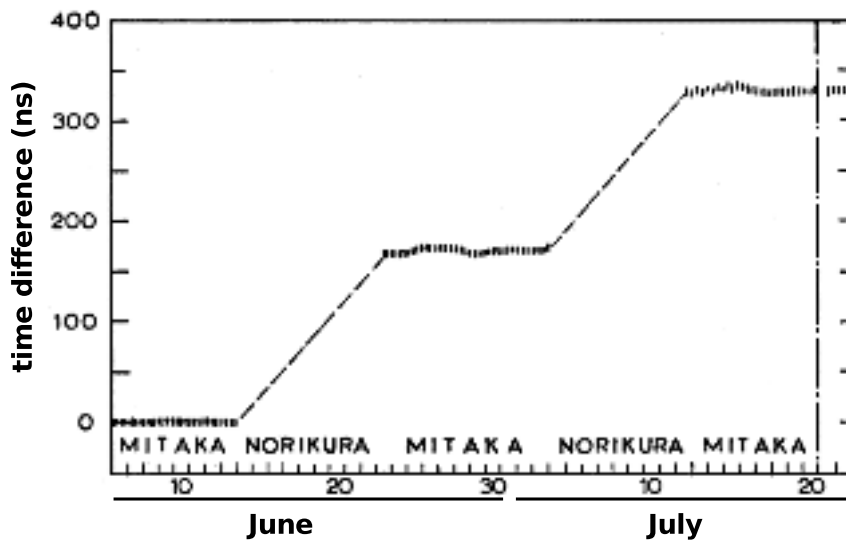
5.2.5 A varying metric

In the Pound-Rebka experiment, the nuclei emitting the gamma rays at frequency f can be thought of as little clocks. Each wave crest that propagates upward is a signal saying that the clock has ticked once. An observer at the top of the tower finds that the signals come in at the lower frequency f' , and concludes naturally that the clocks at the bottom had been slowed down due to some kind of time dilation effect arising from gravity.

This may seem like a big conceptual leap, but it has been confirmed using atomic clocks. In a 1978 experiment by Iijima and Fujiwara, figure h, identical atomic clocks were kept at rest at the top and bottom of a mountain near Tokyo. The discrepancies between the clocks were consistent with the predictions of the equivalence principle. The gravitational Doppler shift was also one of the effects that led to the non-null result of the Hafele-Keating experiment p. 15, in which atomic clocks were flown around the world aboard commercial passenger jets. Every time you use the GPS system, you are making use of these effects.

Starting from only the seemingly innocuous assumption of the equivalence principle, we are led to surprisingly far-reaching conclusions. We find that time flows at different rates depending on the height within a gravitational field. Since the metric can be interpreted as a measure of the amount of proper time along a given world-line, we conclude that we cannot always express the metric in the familiar form $\tau^2 = (+1)\Delta t^2 + (-1)\Delta x^2$ with fixed coefficients $+1$ and -1 . Suppose that the t coordinate is defined by radio synchronization. Then the $+1$ in the metric needs to be replaced with

²Phys. Rev. Lett. 4 (1960) 337



h / A graph showing the time difference between two atomic clocks. One clock was kept at Mitaka Observatory, at 58 m above sea level. The other was moved back and forth to a second observatory, Norikura Corona Station, at the peak of the Norikura volcano, 2876 m above sea level. The plateaus on the graph are data from the periods when the clocks were compared side by side at Mitaka. The difference between one plateau and the next shows a gravitational effect on the rate of flow of time, accumulated during the period when the mobile clock was at the top of Norikura.

approximately $1 + 2\Phi$, where we take $\Phi = 0$ by convention at the height of the standard clock that coordinates the synchronization.

Keep in mind that although we have connected gravity to the measurement apparatus of special relativity, there is no curvature of spacetime, so what we are doing here is still special relativity, not general relativity. In fact there is nothing more mysterious going on here than a *renaming* of spacetime events through a change of coordinates. The renaming might be convenient if we were using earth-based reference points to measure the x coordinate. But if we felt like it, we could switch to a good inertial frame of reference, one that was free-falling. In this frame, we would obtain *exactly the same prediction* for the results of any experiment. For example, the free-falling observer would explain the result of the Pound-Rebka experiment as arising from the upward acceleration of the detector away from the source.

Problems

1 Carry out the details of the calculation of the gravitational Doppler effect in section 5.2.4.

2 A student argues as follows. At the center of the earth, there is zero gravity by symmetry. Therefore time would flow at the same rate there as at a large distance from the earth, where there is also zero gravity. Although we can't actually send an atomic clock to the center of the earth, interpolating between the surface and the center shows that a clock at the bottom of a mineshaft would run faster than one on the earth's surface. Find the mistake in this argument.

3 Somewhere in outer space, suppose there is an astronomical body that is a sphere consisting of solid lead. Assume the Newtonian expression $\Phi = -GM/r$ for the potential in the space outside the object. Make an order of magnitude estimate of the diameter it must have if the gravitational time dilation at its surface is to be a factor of 2 relative to time as measured far away. (Under these conditions of strong gravitational fields, special relativity is only a crude approximation, and that's why we won't get more than an order of magnitude estimate out of this.) What is the gravitational field at its surface? If I have a week's vacation from work, and I spend it lounging on the beach on the lead planet, do I experience two weeks of relaxation, or half a week?

Chapter 6

Waves

This chapter and the preceding one have good, solid physical titles. Inertia. Waves. But underlying the physical content is a thread of mathematics designed to teach you a language for describing space-time. Without this language, the complications of relativity rapidly build up and become unmanageable. In section 5.2.5, we saw that there are physically compelling reasons for switching back and forth between different coordinate systems — different ways of attaching *names* to the events that make up spacetime. A toddler in a bilingual family gets a payoff for switching back and forth between asking *Mamá* in Spanish for *dulces* and alerting Daddy in English that Barbie needs to be rescued from falling off the couch. She may bounce back and forth between the two languages in a single sentence — a habit that linguists call “code switching.” In relativity, we need to build fluency in a language that lets us talk about actual phenomena without getting hung up on the naming system.

6.1 Frequency

6.1.1 Is time’s flow constant?

The simplest naming task is in $0 + 1$ dimensions: a time-line like the ones in history class. If we name the points in time A, B, C, ... or 1, 2, 3, ..., or Bush, Clinton, Bush, ..., how do we know that we’re marking off equal time intervals? Does it make sense to imagine that time itself might speed up and slow down, or even start and stop? The second law of thermodynamics encourages us to think that it could. If the universe had existed for an infinite time, then entropy would have maximized itself — a long time ago, presumably — and we would not exist, because the heat death of the universe would already have happened.

6.1.2 Clock-comparison experiments

But what would it actually mean empirically for time’s rate of flow to vary? Unless we can tie this to the results of experiments, it’s nothing but cut-rate metaphysics. In a Hollywood movie where time could stop, the scriptwriters would show us the stopping through the eyes of an observer, who would stroll past frozen waterfalls and snapshotted bullets in mid-flight. The observer’s brain is a kind of clock, and so is the waterfall. We’re left with what’s known as a clock-comparison experiment. To date, all clock-comparison

experiments have given null results. Matsakis *et al.*¹ found that pulsars match the rates of atomic clocks with a drift of less than about 10^{-6} seconds over 10 years. Guéna *et al.*² observed that atomic clocks using atoms of different isotopes drifted relative to one another by no more than about 10^{-16} per year. Any non-null result would have caused serious problems for relativity. One of the expectations in an Aristotelian description of spacetime is that the motion of material objects on earth would naturally slow down relative to celestial phenomena such as the rising and setting of the sun. The relativistic interpretation of time dilation as an effect on time itself (p. 26) also depends crucially on the null results of these experiments.

6.1.3 Birdtracks notation

As a simple example of clock comparison, let's imagine using the hourly emergence of a mechanical bird from a pendulum-driven cuckoo clock to measure the rate at which the earth spins. There is clearly a kind of symmetry here, since we could equally well take our planet's rotation as the standard and use it to measure the frequency with which the bird pops out of the door. Schematically, let's represent this measurement process with the following notation, which is part of a system called birdtracks:³

$$c \rightarrow e = 24$$

Here c represents the cuckoo clock and e the rotation of the earth. Although the measurement relationship is nearly symmetric, the arrow has a direction, because, for example, the measurement of the earth's rotational period in terms of the clock's frequency is $c \rightarrow e = (1 \text{ hr}^{-1})(24 \text{ hr}) = 24$, but the clock's period in terms of the earth's frequency is $e \rightarrow c = 1/24$. We say that the relationship is not symmetric but "dual." By the way, it doesn't matter how we arrange these diagrams on the page. The notations $c \rightarrow e$ and $e \leftarrow c$ mean exactly the same thing, and expressions like this can even be drawn vertically.

Suppose that e is a displacement along some one-dimensional line of time, and we want to think of it as the thing being measured. Then we expect that the measurement process represented by c produces a real-valued result and is a linear function of e . Since the relationship between c and e is dual, we expect that c also belongs to some vector space. For example, vector spaces allow multiplication by a scalar: we could double the frequency of the cuckoo clock

¹Astronomy and Astrophysics 326 (1997) 924,
adsabs.harvard.edu/full/1997A&26A...326..924M

²arxiv.org/abs/1205.4235

³The system used in this book follows the one defined by Cvitanović, which was based closely on a graphical notation due to Penrose. For a more complete exposition, see the Wikipedia article "Penrose graphical notation" and Cvitanović's online book at birdtracks.eu.

by making the bird come out on the half hour as well as on the hour, forming $2c$. Measurement should be a linear function of both vectors; we say it is “bilinear.”

6.1.4 Duality

The two vectors c and e have different units, hr^{-1} and hr , and inhabit two different one-dimensional vector spaces. The “flavor” of the vector is represented by whether the arrow goes into it or comes out. Just as we used notation like \vec{v} in freshman physics to tell vectors apart from scalars, we can employ arrows in the birdtracks notation as part of the notation for the vector, so that instead of writing the two vectors as c and e , we can notate them as $c \rightarrow$ and $\rightarrow e$. Performing a measurement is like plumbing. We join the two “pipes” in $c \rightarrow \rightarrow e$ and simplify to $c \rightarrow e$.

A confusing and nonstandardized jungle of notation and terminology has grown up around these concepts. For now, let’s refer to a vector such as $\rightarrow e$, with the arrow coming in, simply as a “vector,” and the type like $c \rightarrow$ as a “covector.” In the one-dimensional example of the earth and the cuckoo clock, the roles played by the two things were completely equivalent, and it didn’t matter which one we expressed as a vector and which as a covector.

6.2 Phase

6.2.1 Phase is a scalar

In section 1.3.1, p. 22, we defined a (Lorentz) *invariant* as a quantity that was unchanged under rotations and Lorentz boosts. A measurement such as $c \rightarrow e = 24$ is an invariant because it is simply a count. We’ve counted the number of periods. In fact, a count is not just invariant under rotations and boosts but under any well-behaved change of coordinates — the technical condition being that each coordinate in each set is a differentiable function of each coordinate in the other set. Such a change of coordinates is called a *diffeomorphism*. For example, a uniform scaling of the coordinates $(t, x, y, z) \rightarrow (kt, kx, ky, kz)$, which is analogous to a change of units,⁴ is all right as long as k is nonzero. A quantity that stays the same under any diffeomorphism is called a *scalar*. Since a Lorentz transformation is a diffeomorphism, every scalar is a Lorentz invariant. Not every Lorentz invariant is a scalar.

The determinant of the metric

Example 1

Minkowski coordinates can be defined as coordinates in which the metric has the standard form $g = \text{diag}(1, -1, -1, -1)$. If we rescale these coordinates according to $(t, x, y, z) \rightarrow (kt, kx, ky, kz)$, then the metric changes according to $g \rightarrow k^{-2}g$. To keep track of

⁴The appropriate relativistic way of defining a change of units is subject to some ambiguity. See section 9.6, p. 207.

how “un-Minkowski” this scaling is, we could use the determinant of the metric $\det(g) = -k^{-8}$. This determinant tells us how many coordinate-grid boxes fit in a unit volume, and it is of interest in a more general context than this example of uniform rescaling, e.g., it serves a similar function when converting from Cartesian coordinates to polar coordinates.

Under a Lorentz transformation or a rotation, the metric retains its standard form, and therefore $\det(g)$ is Lorentz invariant. Another way of seeing this is that spacetime volume is Lorentz invariant (p. 49), so that a Lorentz transformation doesn’t change how many coordinate-grid boxes fit in a unit volume.

But although $\det(g)$ is a Lorentz invariant, it is not a scalar, because it changes under the transformation described above.

In birdtracks notation, any expression that has no external arrows at all represents a scalar. Since the expression $c \rightarrow e = 24$ has no external arrows, only internal ones, it represents a scalar. Another way of describing this measurement is as a phase. If we prefer to measure the phase ϕ in units of cycles, then we have $\phi = c \rightarrow e$. If we like radians, we can use $\phi = 2\pi c \rightarrow e$.

6.2.2 Scaling

A convenient way of summarizing all of our categories of variables is by their behavior when we rescale our coordinates. If we switch our time unit from hours to minutes, the number of apples in a bowl is unchanged, the earth’s period of rotation gets 60 times bigger, and the frequency of the cuckoo clock changes by a factor of $1/60$. In other words, a quantity u under rescaling of coordinates by a factor α becomes $\alpha^p u$, where the exponents -1 , 0 , and $+1$ correspond to covectors, scalars, and vectors, respectively. We can therefore see that these distinctions are of interest even in one dimension, contrary to what one would have expected from the freshman-physics concept of a vector as something transforming in a certain way under rotations.

In section 1.3.1 (p. 22), we defined an invariant as a quantity that did not change under rotations or Lorentz boosts, i.e., one that was independent of the frame of reference. For a scalar we have the even more restrictive condition that it must not change under any change of coordinates. For example, area in $1 + 1$ -dimensional spacetime is an invariant, but it’s not a scalar; it changes when we rescale our coordinates.

6.3 The frequency-wavenumber covector

Generalizing from $0 + 1$ dimensions to $3 + 1$, we could have an observer moving inertially along velocity vector $\rightarrow \mathbf{o}$, while counting the phase ϕ (in radians) of a plane wave (perhaps a water wave or

an electromagnetic wave) that is washing over her. Since ϕ is just a count, it's clearly a scalar. That means that we have some function that takes as its input a vector $\rightarrow \mathbf{o}$ and gives as an output the scalar ϕ . This function has all the right characteristics to be described as a measurement $\omega \rightarrow \mathbf{o}$ of $\rightarrow \mathbf{o}$ with some covector $\omega \rightarrow$, and in a constructive style of mathematics this is a good way of *defining* a covector: it's a linear function from the space of vectors to the real numbers. We call $\omega \rightarrow$ the frequency-wavelength covector, or just the frequency covector for short. If $\rightarrow \mathbf{o}$ represents one second as measured on the clock of this observer, then $\omega \rightarrow \mathbf{o}$ is the frequency ω measured by this observer in units of radians per second. If the same observer considers \mathbf{s} to be a vector of simultaneity with a length of one meter, then $\omega \rightarrow \mathbf{s}$ is the observer's measurement of the wavenumber k , defined as 2π divided by the wavelength.

6.3.1 Visualization

In more than one dimension, there are natural ways of visualizing the different vector spaces inhabited by vectors and covectors. A vector is an arrow. A covector can be visualized as a set of parallel, evenly spaced lines on a topographic map, a/2, with an arrowhead to show which way is "uphill." The act of measurement consists of counting how many of these lines are crossed by a certain vector, a/3.

Parallelism between vectors and covectors

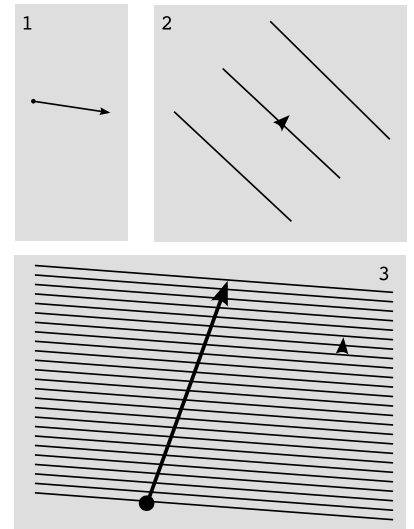
Example 2

It seems visually obvious in figure a/3 that the vector and covector are almost, but not exactly, parallel, since the arrowheads point in almost the same direction. Ordinarily, parallelism of nonzero vectors \mathbf{u} and \mathbf{v} would be expressed by the existence of a real number α such that $\mathbf{u} = \alpha \mathbf{v}$. But vectors and covectors are different kinds of beasts, belonging to different vector spaces. Scaling up a zebra will never produce a giraffe. If there is no metric, then this is simply a fact of life: there is no natural way to define parallelism between a vector \mathbf{v} and a covector ω .

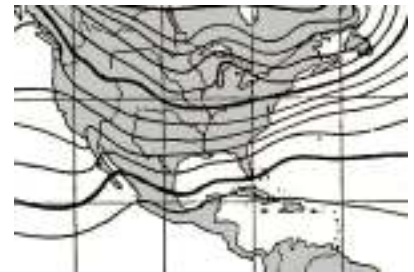
But if we have a metric, then we can define a magnitude for the vector \mathbf{c} in a/3, and keep that magnitude constant while rotating \mathbf{v} . If the metric is Euclidean, then this corresponds to rigidly rotating the arrow on the page, and $\omega \rightarrow \mathbf{v}$ is maximized for a certain orientation, which we define as the condition for parallelism. If the metric is noneuclidean, then things get a bit more complicated, but the same ideas apply if the vectors are either both spacelike or both timelike. For example, if both are timelike, then $\omega \rightarrow \mathbf{v}$ is minimized by parallelism, because the Cauchy-Schwarz inequality is reversed (see sec. 1.5.1, p. 36.)

6.3.2 The gradient

Given a scalar field ϕ , its gradient $\nabla \phi$ at any given point is a covector. The frequency covector is the gradient of the phase. In



a / 1. A displacement vector. 2. A covector. 3. Measurement is reduced to counting. The observer, represented by the displacement vector, counts 24 wavefronts.



b / Constant-temperature curves for January in North America, at intervals of 4°C . The temperature gradient at a given point is a covector.

birdtracks notation, we indicate this by writing it with an outward-pointing arrow, $(\nabla\phi)\rightarrow$. Because gradients occur so frequently, birdtracks notation has a special shorthand for them, which is simply a circle:



This notation can also be extended to the case where the thing being differentiated is not a scalar, but then some complications are encountered when the coordinates are not Minkowski; see section 9.4.1, p. 196.

Cosmological observers

Example 3

Time is relative, so what do people mean when they say that the universe is 13.8 billion years old? If a hypothetical observer had been around since shortly after the big bang, the time elapsed on that observer's clock would depend on the observer's world line. Two such observers, who had had different world-lines, could have differing clock readings.

Modern cosmologists aren't naive about time dilation. They have in mind a cosmologically preferred world-line for their observer. One way of constructing this world-line is as follows. Over time, the temperature T of the universe has decreased. (We define this temperature locally, but we average over large enough regions so that local variations don't matter.) The negative gradient of this temperature, $-\nabla T$, is a covector that points in a preferred direction in spacetime, and a preferred world-line for an observer is one whose velocity vector \mathbf{v} is always parallel to $-\nabla T$, in the sense defined in example 2 above.

6.4 Duality

6.4.1 Duality in 3+1 dimensions

In our original 0 + 1-dimensional example of the cuckoo clock and the earth, we had duality: the measurements $c\rightarrow e = 24$ and $e\rightarrow c = 1/24$ really provided the same information, and it didn't matter whether we made our scalar out of covector $c\rightarrow$ and vector $\rightarrow e$ or covector $e\rightarrow$ and vector $\rightarrow c$. All these quantities were simply clock rates, which could be described either by their frequencies (covectors) or their periods (vectors).

To generalize this to 3 + 1 dimensions, we need to use the metric — a piece of machinery that we have never had to employ since the beginning of the chapter. Given a vector $\rightarrow \mathbf{r}$, suppose we knew how to produce its covector version $\mathbf{r}\rightarrow$. Then we could hook up the plumbing to form $\mathbf{r}\rightarrow \mathbf{r}$, which is just a number. What number could it be? The only reasonable possibility is the squared magnitude of \mathbf{r} , which we calculate using the metric as $r^2 = g(\mathbf{r}, \mathbf{r})$. Since we can think of covectors as functions that take vectors to real numbers, clearly $\mathbf{r}\rightarrow$ should be the function f defined by $f(\mathbf{x}) = g(\mathbf{r}, \mathbf{x})$.

Finding the dual of a given vector**Example 4**

▷ Given the vector $\rightarrow \mathbf{v} = (3, 4)$ in 1 + 1-dimensional Minkowski coordinates, find the covector $\mathbf{v} \rightarrow$, i.e., its dual.

▷ Our goal is to write out an explicit expression for the covector in component form,

$$\mathbf{v} \rightarrow = (a, b).$$

To define these components, we have to have some basis in mind, consisting of one timelike observer-vector \mathbf{o} and one space-like vector of simultaneity \mathbf{s} . Since we're doing this in Minkowski coordinates (section 1.2, 20), let's notate these as $\rightarrow \hat{\mathbf{t}}$ and $\rightarrow \hat{\mathbf{x}}$, where the hats indicate that these are unit vectors in the sense that $\hat{t}^2 = 1$ and $\hat{x}^2 = -1$. Writing $\mathbf{v} \rightarrow$ in terms of a and b means that we're identifying $\mathbf{v} \rightarrow$ with the function f defined by $f(\mathbf{x}) = g(\rightarrow \mathbf{v}, \mathbf{x})$. Therefore

$$f(\rightarrow \hat{\mathbf{t}}) = a \quad \text{and} \quad f(\rightarrow \hat{\mathbf{x}}) = b$$

or

$$g(\rightarrow \mathbf{v}, \rightarrow \hat{\mathbf{t}}) = 3 = a \quad \text{and} \quad g(\rightarrow \mathbf{v}, \rightarrow \hat{\mathbf{x}}) = -4 = b.$$

The result of the formidable, fancy-looking calculation in example 4 was simply to take the vector

$$(3, 4)$$

and flip the sign of its spacelike component to give the its dual, the covector

$$(3, -4).$$

Looking back at why this happened, it was because we were using Minkowski coordinates, and in Minkowski coordinates the form of the metric is $g(p, q) = (+1)p_t q_t + (-1)p_x q_x + \dots$. Therefore, we can always find duals in this way, provided that (1) we're using Minkowski coordinates, and (2) the signature of the metric is, as assumed throughout this book, + - - -, not - + + +.

Going both ways**Example 5**

▷ Assume Minkowski coordinates and signature + - - -. Given the vector

$$\rightarrow \mathbf{e} = (8, 7)$$

and the covector

$$\mathbf{f} \rightarrow = (1, 2),$$

find $\mathbf{e} \rightarrow$ and $\rightarrow \mathbf{f}$

▷ By the rule established above, we can find $\mathbf{e} \rightarrow$ simply by flipping the sign of the 7,

$$\mathbf{e} \rightarrow = (8, -7).$$

To find $\rightarrow \mathbf{f}$, we need to ask what vector (a, b) , if we flipped the sign of b , would give us $(a, -b) = (1, 2)$. Obviously this is

$$\rightarrow \mathbf{f} = (1, -2).$$

In other words, flipping the sign of the spacelike part of a vector is also the recipe for changing covectors into vectors.

Example 5 shows that in Minkowski coordinates, the operation of changing a covector to the corresponding vector is the *same* as that of changing a vector to its covector. Thus, the dual of a dual is the same thing you started with. In this respect, duality is similar to arithmetic operations such as $x \rightarrow -x$ and $x \rightarrow 1/x$. That is, the duality is a self-inverse operation — it undoes itself, like getting two sex-change operations in a row, or switching political parties twice in a country that has a two-party system. Birdtracks notation makes this self-inverse property look obvious, since duality means switching a inward arrow to an outward one or vice versa, and clearly doing two such switches gives back the original notation. This property was established in example 5 by using Minkowski coordinates and assuming the signature to be $+ - - -$, but it holds without these assumptions (problem 1, p. 142).

In the general case where the coordinates may not be Minkowski, the above analysis plays out as follows. Covectors and vectors are represented by row and column vectors. The metric can be specified by a matrix g so that the inner product of column vectors p and q is given by $p^T g q$, where T represents the transpose. Rerunning the same logic with these additional complications, we find that the dual of a vector q is $(gq)^T$, while the dual of a covector ω is $(\omega g^{-1})^T$, where g^{-1} is the inverse of the matrix g .

6.4.2 Change of basis

We saw in section 6.2.2 that in $0 + 1$ dimensions, vectors and covectors has opposite scaling properties under a change of units, so that switching our base unit from hours to minutes caused our frequency covectors to go up by a factor of 60, while our time vectors went down by the same factor. This behavior was necessary in order to keep scalar products the same. In more than one dimension, the notion of changing units is replaced with that of a change of basis. In linear algebra, row vectors and column vectors act like covectors and vectors; they are dual to each other. Let B be a matrix made of column vectors, representing a basis for the column-vector space. Then a change of basis for a row vector r is expressed as $r' = rB$, while the same change of basis for a column vector c is $c' = B^{-1}c$. We then find that the scalar product is unaffected by the change of basis, since $r'c' = rBB^{-1}c = rc$.

In the important special case where B is a Lorentz transformation, this means that covectors transform under the *inverse* trans-

formation, which can be found by flipping the sign of v . This fact will be important in the following section.

6.5 The Doppler shift and aberration

6.5.1 Doppler shift

As an example, we generalize our previous discussion of the Doppler shift of light to $3 + 1$ dimensions.

For clarity, let's first show how the $1 + 1$ -dimensional case works in our new notation. For a wave traveling to the left, we have $\omega \rightarrow = (\omega, \omega)$ (not $(\omega, -\omega)$ — see figure d/1). We now want to transform into the frame of an observer moving to the right with velocity v relative to the original frame. Because $\omega \rightarrow$ is a covector, we do this using the *inverse* Lorentz transformation. An ordinary Lorentz transformation would take a lightlike vector (ω, ω) to $(\omega/D, \omega/D)$ (see section 3.2). The inverse Lorentz transformation gives $(D\omega, D\omega)$. The frequency has been shifted upward by the factor D , as established previously.

In $3 + 1$ dimensions, a spatial plane is determined by the light's direction of propagation and the relative velocity of the source and observer, so this case reduces without loss of generality to $2 + 1$ dimensions. The frequency four-vector must be lightlike, so its most general possible form is $(\omega, \omega \cos \theta, \omega \sin \theta)$, where θ is interpreted as the angle between the direction of propagation and the relative velocity. In $2 + 1$ dimensions, a Lorentz boost along the x axis looks like this:

$$\begin{aligned} t' &= \gamma t - v\gamma x \\ x' &= -v\gamma t + \gamma x \\ y' &= y \end{aligned}$$

The inverse transformation is found by flipping the sign of v . Putting our frequency vector through an inverse Lorentz boost, we find

$$\omega' = \gamma\omega(1 + v \cos \theta).$$

For $\theta = 0$ the Doppler factor reduces to $\gamma(1 + v) = D$, recovering the $1 + 1$ -dimensional result. For $\theta = 90^\circ$, we have $\omega' = \gamma\omega$, which is interpreted as a pure time dilation effect when the source's motion is transverse to the line of sight.

To see the power of the mathematical tools we've developed in this chapter, you may wish to look at sections 6 and 7 of Einstein's 1905 paper on special relativity, where a lengthy derivation is needed in order to arrive at the same result.

6.5.2 Aberration

Imagine that rain is falling vertically while you drive in a convertible with the top down. To you, the raindrops appear to be

moving at some nonzero angle relative to vertical. This is referred to as aberration: a world-line's direction changes depending on one's frame of reference. In the street's frame of reference, the angle between the rain's three-velocity and the car's is $\theta = 90^\circ$, but in the car's frame $\theta' \neq 90^\circ$. In this example, aberration is a large effect because the car's speed v is comparable to the velocity u of the raindrops. To a snail crawling along the sidewalk at a much lower v , the effect would be small. Using the small-angle approximation $\tan \epsilon \approx \epsilon$, we find that for small v , the difference $\Delta\theta = \theta' - \theta$ would be approximately v/u , in units of radians.

Compared to a ray of light, we're all like snails. For example, the earth's orbital speed is about $v \sim 10^{-4}$ in units where the speed of light $u = 1$, so we expect a maximum effect of about 10^{-4} radians, or $20''$ of arc, which is small but not negligible for a telescope with a high-quality mount, being used at high magnification.

This estimate of astronomical aberration of light is roughly right, but we don't expect it to be exact, both because of the small-angle approximation and because we calculated it using a Galilean picture of spacetime. Let's calculate the exact result. As shown in example 8 on p. 137, the direction of propagation of a light wave lies along the vector that is the dual to its frequency covector. Let's call this direction of propagation $\rightarrow \mathbf{u}$. Reusing the expression for $\omega \rightarrow$ defined in section 6.5.1, and arbitrarily fixing $\rightarrow \mathbf{u}$'s timelike component to be 1, we have

$$\rightarrow \mathbf{u} = (1, -\cos \theta, -\sin \theta).$$

When this vector undergoes a boost v along the x axis it becomes

$$\rightarrow \mathbf{u}' = (\gamma(1 + v \cos \theta), \gamma(-v - \cos \theta), -\sin \theta)$$

The original angle $\theta = \tan^{-1}(u_y/u_x)$ has been transformed to $\theta' = \tan^{-1}(u'_y/u'_x)$, the result being

$$\tan \theta' = \frac{\sin \theta}{\gamma(\cos \theta + v)}.$$

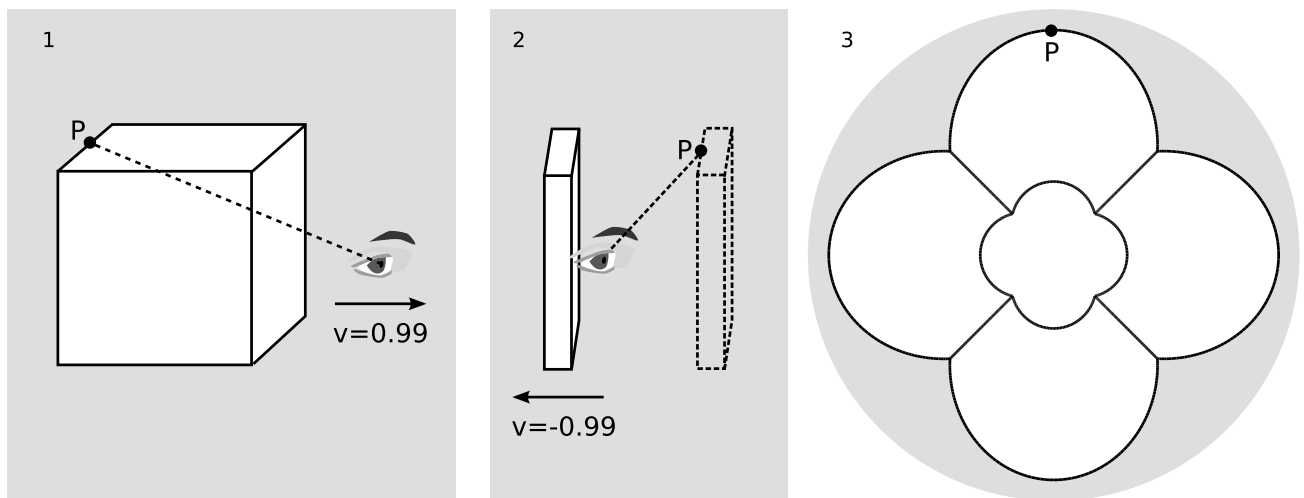
A test of special relativity

Example 6

An assumption underlying this treatment of aberration was that the speed of light was $u = c$, regardless of the velocity of the source. Not all prerelativistic theories had this property, and one would expect that in such a theory, aberration would not be in accord with the relativistic result. In particular, suppose that we believed in Galilean spacetime, so that when a distant galaxy, receding from us at some speed w , emitted a ray of light toward us, the light's velocity in our frame was $u = c - w$. That is, we imagine a theory in which emitting a ray of light is like shooting a bullet from a gun. Since aberration effects go approximately like v/u , we would expect that the reduced u would lead to more aberration compared to the prediction of relativity.

To test theories of this type, Heckmann⁵ used a 24-inch reflector at Hamburg to take high-magnification photographic plates of a star field in Ursa Major containing 11 stars inside the Milky Way and 5 distant galaxies. Measurements of Doppler shifts showed that the galaxies were receding from us at velocities of about $w = 0.05c$, whereas stars within the Milky Way move relative to us at speeds that are negligible in comparison. If, contrary to the relativistic prediction, this led to a 5% decrease in u , then we would expect about a 5% increase in aberration for the galaxies compared to the stars.

Over the course of a year, the earth's orbit carries it toward and away from Ursa Major, so that in the earth's frame of reference, the stars and galaxies have varying velocities relative to us, and the $\sim 20''$ aberration effect oscillates in direction. If the effect was different for the galaxies and the stars, then they ought to shift their apparent positions relative to one another. The shift ought to be on the order of 5% of $20''$, or one second of arc. The results from the observations showed that these relative positions did not appear to vary at all over the course of a year, with the average relative shift being $0.00 \pm 0.06''$ of arc. This difference in aberration is consistent with zero, as predicted by special relativity.



c / 1. The cube's rest frame. 2. The observer's frame. 3. The observer's view of the cube, severely distorted by aberration.

The view of an ultrarelativistic observer *Example 7*

Figure c shows a visualization for an observer flying through a cube at $v = 0.99$. In c/1, the cube is shown in its own rest frame, where it has sides of unit length, and the observer, having already

⁵Annales d'Astrophysique 23 (1960) 410, adsabs.harvard.edu/abs/1960AnAp...23..410H.

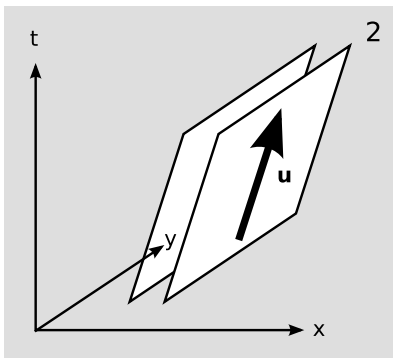
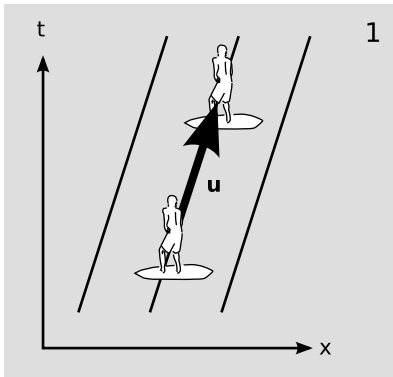
passed through, lies one unit to the right of the cube's center. The observer is facing to the right, away from the cube. The dashed line is a ray of light that travels from point P to the observer, and in this frame it appears as though the ray, arriving from $\theta = 162^\circ$, would not make it into the observer's eye.

But in the observer's frame, $c/2$, the ray is at $\theta' = 47^\circ$, so it actually does fall within her field of view. The cube is length-contracted by a factor $\gamma \approx 7$. The ray was emitted earlier, when the cube was out in front of the observer, at the position shown by the dashed outline.

The image seen by the observer is shown in $c/3$. The circular outline defining the field of view represents $\theta' = 50^\circ$. Note that the relativistic length contraction is not at all what an observer sees optically. The optical observation is influenced by length contraction, but also by aberration and by the time it takes for light to propagate to the observer. The time of propagation is different for different parts of the cube, so in the observer's frame, $c/2$, rays from different points had to be emitted when the cube was at different points in its motion, if those rays were to reach the eye.

A group at Australian National University has produced animations of similar scenes, which can be found online by searching for “optical effects of special relativity.”

It's fun to imagine the view of an observer aboard an ultrarelativistic starship. For v sufficiently close to 1, any angle $\theta < 180^\circ$ transforms to a small θ' . Thus, all light coming to this observer from the surrounding stars — even those in extreme backward directions! — is gathered into a small, bright patch of light that appears to come from straight ahead. Some visible light would be shifted into the extreme ultraviolet and infrared, while some infrared and ultraviolet light would become visible.



d / The surfer moves directly to the right with velocity vector \mathbf{u} . The wave also propagates to the right.

6.6 Phase and group velocity

6.6.1 Phase velocity

A wavefront is a line or surface of constant phase. In a snapshot of a wave at one moment of time, the direction of propagation of the wave is across the wavefronts. The visual situation is different in a spacetime diagram. In $1 + 1$ dimensions, figure d/1, suppose that the lines represent the crest of the water waves. The surfer is on top of a crest, riding along with it. His velocity vector $\rightarrow \mathbf{u}$ is in the spacetime direction that lies on top of the wavefront, not across it. Clearly both his motion and the propagation of the wave are to the right, not to the left as we might imagine based on experience with snapshots of waves.

In $2 + 1$ dimensions, $d/2$, the surfer's velocity is visualized as an arrow lying within a plane of constant phase. Given the wave's phase information, there is more than one possible arrow of this kind. We could try to resolve the ambiguity by requiring that the arrow's projection into the xy plane be perpendicular to the intersection of the wavefronts with that plane, but (with the exception of the case where the wave travels at c , example 8, p. 137) this prescription gives results that change depending on our frame of reference, and the changes are not describable by a Lorentz transformation of the velocity vector. This shows that in the general case, the phase information of the wave, encoded in the frequency covector $\omega \rightarrow$, does not describe the direction of the wave's propagation through space. At most it tells us the wave's *phase velocity*, ω/k , which is not really a velocity. All of these are symptoms of the fact that a velocity is supposed to be a vector, but $\omega \rightarrow$ is a covector. The phase velocity lacks physical interest, because it is not the velocity at which any "stuff" moves.

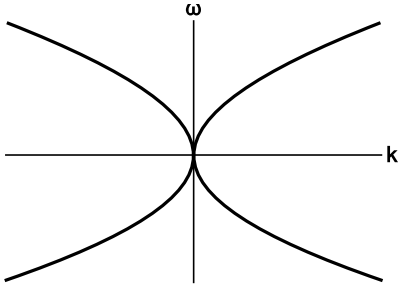
Velocity vector of a light wave, given its phase *Example 8*

We've seen that in general, the information about the phase of a wave encoded in $\omega \rightarrow$ does not determine its direction of propagation. The exception is a wave, such as a light wave, that propagates at c . Let a world-line of propagation of the wave lie along the vector $\rightarrow \mathbf{v}$. In the case of a wave propagating at c , we have $v^2 = 0$ (so that $\rightarrow \mathbf{v}$ can't have the usual normalization for a velocity vector), and the dispersion relation is simply $\omega^2 = 0$. Since the phase stays constant along a world-line of propagation, $\omega \rightarrow \mathbf{v} = 0$. We therefore find that \mathbf{v} and ω are two nonzero, lightlike vectors that are orthogonal to each other. But as shown in problem 10 on p. 39, this implies that the two vectors are parallel. Thus if we're given the covector $\omega \rightarrow$, we just have to compute its dual $\rightarrow \omega$ to find the direction of propagation.

6.6.2 Group velocity

The phase velocity is not the velocity at which "stuff" is transmitted by the wave. The velocity of the stuff is called the group velocity. To have a meaningfully defined group velocity, we need to have a wave that is modulated, because an unmodulated wave is an infinite sine wave that stretches off to infinity, and such an unmodulated wave does not transmit any energy or information. An unmodulated wave has the same frequency covector $\omega \rightarrow$ throughout all of spacetime, i.e., the same frequency ω and wavenumber k . One way of describing a modulated wave is by how $\omega \rightarrow$ does change.

But the different components of $\omega \rightarrow$ are not free to change in any randomly chosen way. Normally they are constrained by a dispersion relation. For example, surface waves in deep water obey the constraint $C = 0$, where $C = \omega^4 - \alpha^2 k^2$ (figure e) and α is a constant with units of acceleration, relating to the acceleration of



e / Points on the graph satisfy the dispersion relation $C = 0$ for water waves. At a given point on the graph, the covector $(\nabla C) \rightarrow$ tells us the group velocity.

gravity. (Since the water is infinitely deep, there is no other scale that could enter into the constraint.)

Now if a certain bump on the envelope with which the wave is modulated visits spacetime events P and Q, then whatever frequency and wavelength the wave has near the bump are observed to be the same at P and Q. In general, k and ω are constant along the spacetime displacement of any point on the envelope, so the spacetime displacement $\rightarrow \mathbf{r}$ from P to Q must satisfy the condition $(\nabla \omega) \rightarrow \mathbf{r} = 0$.

In addition, $\nabla \omega$ must be tangent to the surface of constraint $C = 0$, so that the wave always obeys the constraint. Thus given a point $\omega \rightarrow$ in frequency space, the direction of propagation \mathbf{r} must be uniquely determined by the constraint. Suppose C is a well-behaved function, so that it is approximately a linear function of any small change $\Delta \omega$, i.e., in 1 + 1 dimensions we have

$$\Delta C = \frac{\partial C}{\partial \omega} \Delta \omega + \frac{\partial C}{\partial k} \Delta k.$$

In this approximation, ΔC is a linear function that acts on a covector $\Delta \omega$ and gives back a scalar. In other words, ΔC acts like a vector with components

$$\Delta C = \left(\frac{\partial C}{\partial \omega}, \frac{\partial C}{\partial k} \right).$$

This vector is parallel to \mathbf{r} , so that it points in the wave's direction of propagation through spacetime, and tells us its group velocity $(\partial C / \partial k) / (\partial C / \partial \omega)$. In our example of water waves, a calculation shows that the group velocity is $\pm \alpha / 2\omega$, which is half the phase velocity.

6.7 Abstract index notation

This chapter has centered on the physics of waves, but along the way we've found it helpful to build up some mathematical ideas such as covectors, which have applications in a much broader physical context. In this section we'll develop some related notation.

Expressions in birdtracks notation such as

$$\textcircled{C} \rightarrow s$$

can be awkward to type on a computer, which is why we've already been occasionally resorting to more linear notations such as $(\nabla C) \rightarrow s$. For more complicated birdtracks, the diagrams sometimes look like complicated electrical schematics, and the problem of generating them on a keyboard get more acute. There is in fact a systematic way of representing any such expression using only ordinary subscripts and superscripts. This is called abstract index notation,

and was introduced by Roger Penrose at around the same time he invented birdtracks. For practical reasons, it was the abstract index notation that caught on.

The idea is as follows. Suppose we wanted to describe a complicated birdtrack verbally, so that someone else could draw it. The diagram would be made up of various smaller parts, a typical one looking something like the scalar product $u \rightarrow v$. The verbal instructions might be: “We have an object u with an arrow coming out of it. For reference, let’s label this arrow as a . Now remember that other object v I had you draw before? There was an arrow coming into that one, which we also labeled a . Now connect up the two arrows labeled a .”

Shortening this lengthy description to its bare minimum, Penrose renders it like this: $u_a v^a$. Subscripts depict arrows coming out of a symbol (think of water flowing from a tank out through a pipe below). Superscripts indicate arrows going in. When the same letter is used as both a superscript and a subscript, the two arrows are to be piped together.

Abstract index notation evolved out of an earlier one called the Einstein summation convention, in which superscripts and subscripts referred to specific coordinates. For example, we might take 0 to be the time coordinate, 1 to be x , and so on. A symbol like u_λ would then indicate a component of the dual vector u , which could be its x component if λ took on the value 1. Repeated indices were summed over.

The advantage of the birdtrack and abstract index notations is that they are coordinate-independent, so that an equation written in them is valid regardless of the choice of coordinates. The Einstein and abstract-index notations look very similar, so for example if we want to take a general result expressed in abstract-index notation and apply it in a specific coordinate system, there is essentially no translation required. In fact, the two notations look so similar that we need an explicit way to tell which is which, so that we can tell whether or not a particular result is coordinate-independent. We therefore use the convention that Latin indices represent abstract indices, whereas Greek ones imply a specific coordinate system and can take on numerical values, e.g., $\lambda = 1$.

The following are some examples of equivalent equations written side by side in birdtracks and abstract index notations.

Observer \mathbf{o} ’s displacement in spacetime is a vector:

$$\rightarrow \mathbf{o} \quad o^a$$

In Einstein notation, it’s awkward to express a vector as a whole, because in a notation like o^λ , λ is supposed to take on a particular value. If we used o^λ to mean the whole vector, it would be an abuse

of notation. In abstract index notation, however, the a is simply a name we gave to a pipe coming into vector \mathbf{o} ; the fact that we didn't need to refer to the name in order to connect it to some other pipe is irrelevant.

A wave's frequency is a covector:

$$\omega \rightarrow \quad \omega_a$$

An observer experiences proper time τ :

$$\mathbf{o} \rightarrow \mathbf{o} = \tau^2 \quad o_a o^a = \tau^2$$

There are no external arrows in the birdtracks version, and in the abstract-index version all lower indices (pipes coming out) have been paired with upper indices (pipes coming in); this indicates that the proper time is a scalar, and therefore independent of any choice of coordinate system. In Einstein notation, this becomes $o_\lambda o^\lambda$, with an implied sum over the repeated index, $\sum_\lambda o_\lambda o^\lambda$. The λ refers to a particular coordinate system, so in the Einstein notation it is no longer obvious that the equation holds regardless of our choice of coordinates.

A world-line along which a wave propagates lies along a vector that is orthogonal to the wave's frequency covector:

$$\omega \rightarrow \mathbf{u} = 0 \quad \omega_a u^a = 0$$

The frequency covector is the gradient of the phase:

$$\omega \rightarrow = \left(\phi \right) \rightarrow \quad \omega_a = \nabla_a \phi$$

The following grammatical rules apply to both abstract-index and Einstein notation:

1. Repeated indices occur in pairs, with one up and one down and the two factors multiplying each other.
2. Disregarding indices that are paired as in rule 1, all other indices must appear uniformly in all terms and on both sides of an equation. "Appear uniformly" means that an index can't be missing and can't be a superscript in some places but a subscript in others.
3. For reasons to be explained in section 7.4, p. 148, a partial derivative with respect to a coordinate, such as $\partial/\partial x^k$, is treated as if the index were a *subscript*, and conversely $\partial/\partial x_k$ is considered to have a superscripted k .

In abstract-index notation, rule 1 follows because the indices are simply labels describing how, in birdtracks notation, the pipes should

be hooked up. Violating rule 1, as in an expression like $v^a v^a$, produces a quantity that does not actually behave as a scalar. An example of a violation of rule 2 is $v^a = \omega_a$. This doesn't make sense, for the same reason that it doesn't make sense to equate a row vector to a column vector in linear algebra. Even if an equation like this did hold in one frame of reference, it would fail in another, since the left-hand and right-hand sides transform differently under a boost.

In section 6.4.1 we discussed the notion of finding the covector that was dual to a given vector, and the vector dual to a given covector. Because the distinction between vectors and covectors is represented in index notation by placing the index on the top or on the bottom, relativists refer to this kind of thing as raising and lowering indices. In general, this type of manipulation is called “index gymnastics.” Here's what raising and lowering indices looks like.

Converting a vector to its covector form:

$$u_a = g_{ab} u^b$$

Changing a covector to the corresponding vector:

$$u^a = g^{ab} u_b$$

The symbol g^{ab} refers to the inverse of the matrix g_{ab} .

Problems

1 In section 6.4.1, I proved that duality is a self-inverse operation, invoking Minkowski coordinates and assuming the signature to be $+- - -$. Show that these assumptions were not necessary.

Chapter 7

Coordinates

In your previous study of physics, you’ve seen many examples where one coordinate system makes life easier than another. For a block being pushed up an inclined plane, the most convenient choice may be to tilt the x and y axes. To find the moment of inertia of a disk we use cylindrical coordinates. The same is true in relativity. Minkowski coordinates are not always the most convenient. In chapter 6 we learned to classify physical quantities as covectors, scalars, and vectors, and we learned rules for how these three types of quantities transformed in two special changes of coordinates:

1. When we rescale all coordinates by a factor α , the components of vectors, scalars, and covectors scale by α^p , where $p = +1$, 0 , and -1 , respectively.
2. Under a boost, the three cases require respectively the Lorentz transformation, no transformation, and the inverse Lorentz transformation.

In this chapter we’ll learn how to generalize this to any change of coordinates,¹ and also how to find the form of the metric expressed in non-Minkowski coordinates.

7.1 An example: accelerated coordinates

Let’s start with a concrete example that has some physical interest. In section 5.2, p. 120, we saw that we could have “gravity without gravity,” an experiment carried out in a uniform gravitational field can be interpreted as an experiment in flat spacetime (so that special relativity applies), but with the measurements expressed in the accelerated frame of the earth’s surface. In the Pound-Rebka experiment, all of the results could have been expressed in an inertial (free-falling) frame of reference, using Minkowski coordinates, but this would have been extremely inconvenient, because, for example, they didn’t want to drop their expensive atomic clocks and take the readings before the clocks hit the floor and were destroyed.

Since this is “gravity without gravity,” we don’t actually need a planet cluttering up the picture. Imagine a universe consist-

¹We do require the change of coordinates to be smooth in the sense defined on p. 127, i.e., it should be a diffeomorphism.

ing of limitless, empty, flat spacetime. Describe it initially using Minkowski coordinates (t, x, y, z) . Now suppose we want to find a new set of coordinates (T, X, Y, Z) that correspond to the frame of reference of an observer aboard a spaceship accelerating in the x direction with a constant acceleration.

The Galilean answer would be $X = x - \frac{1}{2}at^2$. But this is unsatisfactory from a relativistic point of view for several reasons. At $t = c/a$ the observer would be moving at the speed of light, but relativity doesn't allow frames of reference moving at c (section 3.4, p. 59). At $t > c/a$, the observer's motion would be faster than c , but this is impossible in $3 + 1$ dimensions (section 3.8, p. 69).

These problems are related to the fact that the observer's *proper* acceleration, i.e., the reading on an accelerometer aboard the ship, isn't constant if $x = \frac{1}{2}at^2$. We saw in example 4 on p. 61 that constant proper acceleration is described by $x = \frac{1}{a} \cosh a\tau$, $t = \frac{1}{a} \sinh a\tau$, where τ is the proper time. For this motion, the velocity only approaches c asymptotically. This suggests the following for the relationship between the two sets of coordinates:

$$t = X \sinh T$$

$$x = X \cosh T$$

$$y = Y$$

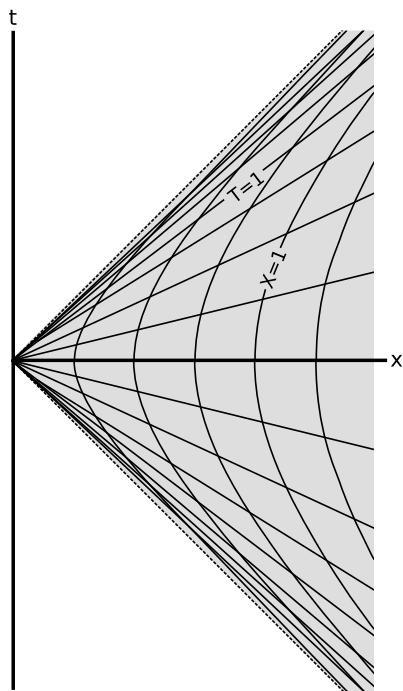
$$z = Z$$

For example, if the ship follows a world-line $(T, X) = (\tau, 1)$, then its motion in the unaccelerated frame is $(t, x) = (\sinh \tau, \cosh \tau)$, which is of the desired form with $a = 1$.

The (T, X, Y, Z) coordinates, called Rindler coordinates, have many, but not all, of the properties we would like for an accelerated frame. Ideally, we'd like to have all of the following: (1) the proper acceleration is constant for any world-line of constant (X, Y, Z) ; (2) the proper acceleration is the same for all such world-lines, i.e., the fictitious gravitational field is *uniform*; and (3) the description of the accelerated frame *is* just a change of coordinates, i.e., we're just talking about the flat spacetime of special relativity, with events renamed. It turns out that we can pick two out of three of these, but it's not possible to satisfy all three at the same time. Rindler coordinates satisfy conditions 1 and 3, but not 2. This is because the proper acceleration of a world-line of constant (X, Y, Z) can easily be shown to be $1/X$, which depends on X . Thus we don't speak of Rindler coordinates as "the" coordinates of an accelerated observer.

Rindler coordinates have the property that if a rod extends along the X axis, and external forces are applied to it in just such a way that every point on the rod has constant X , then it accelerates along its own length without any stress. (See problem 7, p. 214.)

The diagonals are event horizons (p. 62). Their intersection lies along every constant- T line; cf. example 17, p. 34, and p. 73.



a / The transformation between Minkowski coordinates (t, x) and the accelerated coordinates (T, X)

7.2 Transformation of vectors

Now suppose we want to transform a vector whose components are expressed in the (T, X) coordinates into components expressed in (t, x) . Our most basic example of a vector is a displacement $(\Delta T, \Delta X)$, and if we make this an infinitesimal (dT, dX) then we don't need to worry about the fact that the chart in figure a has curves on it — close up, curves look like straight lines.² If we think of the coordinate t as a function of two variables, $t = t(T, X)$, then t is changing for two different reasons: its first input T changes, and also its second input X . If t were only a function of one variable $t(T)$, then the change in t would be given simply by the chain rule, $dt = \frac{dt}{dT} dT$. Since it actually has two such reasons to change, we add the two changes:

$$dt = \frac{\partial t}{\partial T} dT + \frac{\partial t}{\partial X} dX$$

The derivatives are partial derivatives, and these derivatives exist because, as we will always assume, the change of coordinates is smooth. An exactly analogous expression applies for dx .

$$dx = \frac{\partial x}{\partial T} dT + \frac{\partial x}{\partial X} dX$$

Before we carry out the details of this calculation, let's stop and note that the results so far are completely general. Since we have so far made no use of the actual equations for this particular change of coordinates, these expressions would apply to *any* such transformation, including the special cases we've encountered so far, such as Lorentz transformations and scaling. (For example, if we'd been scaling by a factor α , then all of the partial derivatives would simply have equaled α .) Furthermore, our definition of a vector is that a vector is anything that transforms like a vector. Since we've established that the rules above apply to a displacement vector, we conclude that they would also apply to any other vector, say an energy-momentum vector.

Returning to this specific example, application of the facts $d \sinh u / du = \cosh u$ and $d \cosh u / du = \sinh u$ tells us that the vector

$$(dT, dX)$$

is transformed to:

$$(dt, dx) = (X \cosh T dT + \sinh T dX, X \sinh T dT + \cosh T dX)$$

As an example of how this applies universally to any type of vector, suppose that the observer aboard a spaceship with world-line

²Here we make use of the fact that the change of coordinate was smooth, i.e., a diffeomorphism. Otherwise the curves could have kinks in them that would still look like kinks under any magnification.

$(T, X) = (\tau, 1)$ has a favorite paperweight with mass m . According to measurements carried out aboard her ship, its energy-momentum vector is

$$(p_T, p_X) = (m, 0).$$

In the unaccelerated coordinates, this becomes

$$\begin{aligned}(p_t, p_x) &= (X \cosh T p_T + \sinh T p_X, X \sinh T p_T + \cosh T p_X) \\ &= (mX \cosh T, mX \sinh T) \\ &= (m \cosh \tau, m \sinh \tau).\end{aligned}$$

Since the functions \cosh and \sinh behave like e^x for large x , we find that after the astronaut has spent a reasonable amount of proper time τ accelerating, the paperweight's mass-energy and momentum will have grown to the point where it's an awesome weapon of mass destruction, capable of obliterating an entire galaxy.

7.3 Transformation of the metric

Continuing with the example of accelerated coordinates, let's find what happens to the metric when we change from Minkowski coordinates. Minkowski coordinates are essentially defined so that the metric has the familiar form with coefficients $+1$ and -1 . In relativity, one often presents the metric by showing its result when applied to an infinitesimal displacement (dt, dx) :

$$ds^2 = dt^2 - dx^2$$

Here ds would represent proper time, in the case where the displacement was timelike. Since we've already determined that

$$\begin{aligned}dt &= X \cosh T dT + \sinh T dX \quad \text{and} \\ dx &= X \sinh T dT + \cosh T dX,\end{aligned}$$

we can simply substitute into the expression for ds in order to find the form of the metric in (T, X) coordinates. Employing the identity $\cosh^2 - \sinh^2 = 1$, we find

$$ds^2 = X^2 dT^2 - dX^2.$$

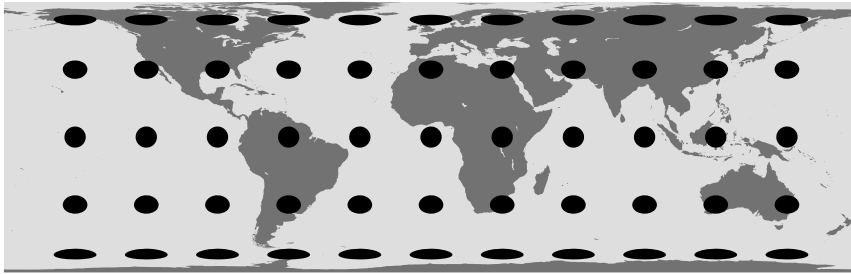
The varying value of the dT^2 coefficient is in fact exactly the kind of gravitational time dilation effect whose existence we predicted in section 5.2.5, p. 122 based on the equivalence principle. The form of the metric inferred there was

$$ds^2 \approx (1 + 2\Delta\Phi) dT^2 - dX^2,$$

where $\Delta\Phi$ is the difference in gravitational potential relative to some reference height. One of the approximations employed was the assumption that the range of heights X was small, but subject to

that approximation, the two results should agree. For convenience, let's consider observers in the region $X \approx 1$, where the acceleration is approximately 1. Then the $\Delta\Phi = \Phi(1 + \Delta X) - \Phi(1) \approx (\text{acceleration})(\text{height}) \approx X$, so the time coefficient in the second form of the metric is $\approx 1 + 2\Delta\Phi \approx 1 + 2\Delta X$. But to within the desired level of approximation, this is the same as $X^2 = (1 + \Delta X)^2 \approx 1 + 2\Delta X$.

The procedure employed above works in general. To transform the metric from coordinates (t, x, y, z) to new coordinates (t', x', y', z') , we obtain the unprimed coordinates in terms of the primed ones, take differentials on both sides, and eliminate t, \dots, dt, \dots in favor of t', \dots, dt', \dots in the expression for ds^2 . We'll see in section 9.2.4, p. 180, that this is an example of a more general transformation law for *tensors*, mathematical objects that generalize vectors and covectors in the same way that matrices generalize row and column vectors. A scalar, with no indices, is called a tensor of rank 0. Vectors and covectors, having one index, are called rank-1 tensors.



b / Example 1.

A map projection

Example 1

Because the earth's surface is curved, it is not possible to represent it on a flat map without distortion. Let ϕ be the latitude, θ the angle measured down from the north pole (known as the colatitude), both measured in radians, and let a be the earth's radius. Then by the definition of radian measure, an infinitesimal north-south displacement by $d\theta$ is a distance $a d\theta$. A point at a given colatitude θ lies at a distance $a \sin \theta$ from the axis, so for an infinitesimal east-west distance we have $a \sin \theta d\phi$. For convenience, let the units be chosen such that $a = 1$. Then the metric, with signature $++$, is

$$ds^2 = d\theta^2 + \sin^2 \theta d\phi.$$

One of the many possible ways of forming a flat map is the Lambert cylindrical projection,

$$\begin{aligned} x &= \phi \\ y &= \cos \theta, \end{aligned}$$

shown in figure b. If we see a distance on the map and want to know how far it actually is on the earth's surface, we need

to transform the metric into the (x, y) coordinates. The inverse coordinate transformation is

$$\begin{aligned}\phi &= x \\ \theta &= \cos^{-1} y.\end{aligned}$$

Taking differentials on both sides, we get

$$\begin{aligned}d\phi &= dx \\ d\theta &= -\frac{dy}{\sqrt{1-y^2}}.\end{aligned}$$

We take the metric and eliminate θ , ϕ , $d\theta$, and $d\phi$, finding

$$ds^2 = (1 - y^2) dx^2 + \frac{1}{1 - y^2} dy^2.$$

In figure b, the polka-dot pattern is made of figures that are actually circles, all of equal size, on the earth's surface. Since they are fairly small, we can approximate y as having a single value for each circle, which means that they are represented on the flat map as approximate ellipses with their east-west dimensions having been stretched by $(1 - y^2)^{-1/2}$ and their north-south ones shrunk by $(1 - y^2)^{1/2}$. Since these two factors are reciprocals of one another, the area of each ellipse is the same as the area of the original circle, and therefore the same as those of all the other ellipses. They are a visual representation of the metric, and they demonstrate the equal-area property of this projection.

7.4 Summary of transformation laws

Having worked through one example in detail, let's progress from the specific to the general. In the Einstein concrete index notation, let coordinates (x^0, x^1, x^2, x^3) be transformed to new coordinates (x'^0, x'^1, x'^2, x'^3) . Then vectors transform according to the rule

$$v'^{\mu} = v^{\kappa} \frac{\partial x'^{\mu}}{\partial x^{\kappa}}, \quad (1)$$

where the Einstein summation convention implies a sum over the repeated index κ . By the same reasoning as in section 6.4.2, p. 132, the transformation for a covector ω is

$$\omega'_{\mu} = \omega_{\kappa} \frac{\partial x^{\kappa}}{\partial x'^{\mu}}. \quad (2)$$

Note the inversion of the partial derivative in one equation compared to the other. Because these equations describe a change from one coordinate system to another, they clearly depend on the coordinate system, so we use Greek indices rather than the Latin ones that would indicate a coordinate-independent abstract index equation.

The letter μ in these equations always appears as an index referring to the new coordinates, κ to the old ones. For this reason, we can get away with dropping the primes and writing, e.g., $v^\mu = v^\kappa \partial x'^\mu / \partial x^\kappa$ rather than v' , counting on context to show that v^μ is the vector expressed in the new coordinates, v^κ in the old ones. This becomes especially natural if we start working in a specific coordinate system where the coordinates have names. For example, if we transform from coordinates (t, x, y, z) to (a, b, c, d) , then it is clear that v^t is expressed in one system and v^c in the other.

In equation (2), μ appears as a subscript on the left side of the equation, but as a superscript on the right. This would appear to violate the grammatical rules given on p. 140, but the interpretation here is that in expressions of the form $\partial/\partial x^i$ and $\partial/\partial x_i$, the superscripts and subscripts should be understood as being turned upside-down. Similarly, (1) appears to have the implied sum over κ written ungrammatically, with both κ 's appearing as superscripts. Normally we only have implied sums in which the index appears once as a superscript and once as a subscript. With our new rule for interpreting indices on the bottom of derivatives, the implied sum is seen to be written correctly. This rule is similar to the one for analyzing the units of derivatives written in Leibniz notation, with, e.g., d^2x/dt^2 having units of meters per second squared. That is, the flipping of the indices like this is required for consistency so that everything will work out properly when we change our units of measurement, causing all our vector components to be rescaled.

The identity transformation

Example 2

In the case of the identity transformation $x'^\mu = x^\mu$, equation (1) clearly gives $v' = v$, since all the mixed partial derivatives $\partial x'^\mu / \partial x^\kappa$ with $\mu \neq \kappa$ are zero, and all the derivatives for $\kappa = \mu$ equal 1.

In equation (2), it is tempting to write

$$\frac{\partial x^\kappa}{\partial x'^\mu} = \frac{1}{\frac{\partial x'^\mu}{\partial x^\kappa}} \quad (\text{wrong!}),$$

but this would give infinite results for the mixed terms! Only in the case of functions of a single variable is it possible to flip derivatives in this way; it doesn't work for partial derivatives. To evaluate these partial derivatives, we have to invert the transformation (which in this example is trivial to accomplish) and then take the partial derivatives.

Polar coordinates

Example 3

None of the techniques discussed here are particular to relativity. For example, consider the transformation from polar coordinates (r, θ) in the plane to Cartesian coordinates

$$\begin{aligned} x &= r \cos \theta \\ y &= r \sin \theta. \end{aligned}$$

A bug sits on the edge of a phonograph turntable, at $(r, \theta) = (1, 0)$. The turntable rotates clockwise, giving the bug a velocity vector $v^\kappa = (v^r, v^\theta) = (0, -1)$, i.e., the angular velocity is one radian per second in the negative (counterclockwise) direction. Let's find the bug's velocity vector in Cartesian coordinates. The transformation law for vectors gives.

$$v^x = v^\kappa \frac{\partial x}{\partial x^\kappa}.$$

Expanding the implied sum over the repeated index κ , we have

$$\begin{aligned} v^x &= v^r \frac{\partial x}{\partial r} + v^\theta \frac{\partial x}{\partial \theta} \\ &= (0) \frac{\partial x}{\partial r} + (-1) \frac{\partial x}{\partial \theta} \\ &= -r \sin \theta \\ &= 0. \end{aligned}$$

For the y component,

$$\begin{aligned} v^y &= v^r \frac{\partial y}{\partial r} + v^\theta \frac{\partial y}{\partial \theta} \\ &= (0) \frac{\partial y}{\partial r} + (-1) \frac{\partial y}{\partial \theta} \\ &= -r \sin \theta \\ &= -1. \end{aligned}$$

7.5 Inertia and rates of change

Suppose that we describe a flying bullet in polar coordinates. We neglect the vertical dimension, so the bullet's motion is linear. If the bullet has a displacement of $(\Delta r_1, \Delta \theta_1)$ in an short time interval Δt , then clearly at a later point in its motion, during an equal interval, it will have a displacement $(\Delta r_2, \Delta \theta_2)$ with two *different* numbers inside the parentheses. This isn't because its velocity or momentum really changed. It's because the coordinate system is curvilinear. There are three ways to get around this:

1. Use only Minkowski coordinates.
2. Instead of characterizing inertial motion as motion with constant velocity components, we can instead characterize it as motion that maximizes the proper time (section 2.4.2, p. 48).
3. Define a correction term to be added when taking the derivative of a vector or covector expressed in non-Minkowski coordinates.

These issues become more acute in general relativity, where curvature of spacetime can make option 1 impossible. Option 3, called the covariant derivative, is discussed in optional section 9.4 on p. 193. If you aren't going to read that section, just keep in mind that in non-Minkowski coordinates, you cannot naively use changes in the components of a vector as a measure of a change in the vector itself.

7.6 ★ Volume, orientation, and the Levi-Civita tensor

This optional section introduces some geometrical machinery that is used in both special and general relativity.

7.6.1 Volume

Desirable properties

In $3 + 1$ dimensions, we have a natural way of defining four-dimensional volume, which is to pick a frame of reference and let the element of volume be $dt dx dy dz$ in the Minkowski coordinates of that frame. Although this definition of 4-volume is stated in terms of certain coordinates, it turns out to be Lorentz-invariant (section 2.5, p. 49). It also has the following desirable properties, which we state for an arbitrary value of m from 1 to 4:

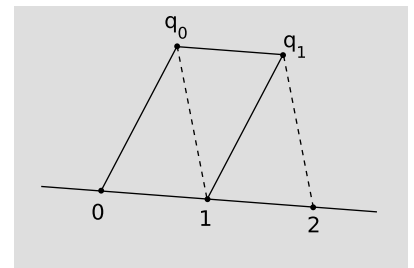
V1. Any two m -volumes can be compared in terms of their ratio.

V2. For any m nonzero vectors, the m -volume of the parallelepiped they span is nonzero if and only if the vectors are linearly independent (that is, if none of them can be expressed in terms of the others using scalar multiplication and vector addition).

We would also like to have convenient methods for working with three-volume, two-volume (area), and one-volume (length). But the m -volumes for $m < 4$ give us headaches if we try to define them so that they obey both V1 and V2. For example, the obvious way to define length ($m = 1$) is to use the metric, but then lightlike vectors would violate V2.

Affine measure

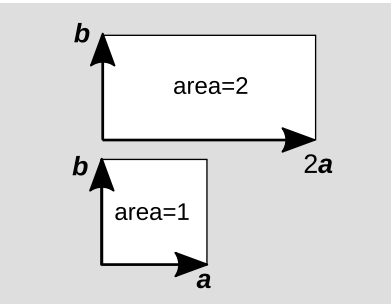
If we're willing to abandon V1, then the following approach succeeds. Consider the $m = 1$ case. We ignore the metric completely and exploit the fact that in special relativity, spacetime is flat (postulate P2, p. 46), so that parallelism works the same way as in Euclidean geometry. Let ℓ be a line, and suppose we want to define a number system on this line that measures how far apart events are. Depending on the type of line, this could be a measurement of time, of spatial distance, or a mixture of the two. First we arbitrarily single out two distinct points on ℓ and label them 0 and 1, as in figure c. Next, pick some auxiliary point q_0 not lying on ℓ . Construct q_0q_1 and parallel to 01 and $1q_1$ parallel to $0q_0$, forming the



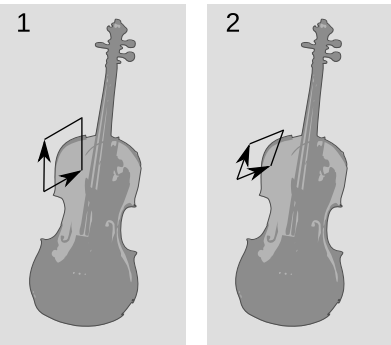
c / Using parallelism to define 1-volume.



d / The area of the viola can be determined by counting the parallelograms formed by the lattice. The area can be determined to any desired precision, by dividing the parallelograms into fractional parts that are as small as necessary.



e / Linearity of area. Doubling the vector **a** doubles the area.



f / The viola has a different area when measured using a different parallelogram as the unit.

parallelogram shown in the figure. Continuing in this way, we have a scaffolding of parallelograms adjacent to the line, determining an infinite lattice of points $1, 2, 3, \dots$ on the line, which represent the positive integers. Fractions can be defined in a similar way. For example, $\frac{1}{2}$ is defined as the point such that when the initial lattice segment $0\frac{1}{2}$ is extended by the same construction, the next point on the lattice is 1. The continuously varying variable constructed in this way is called an *affine parameter*. The time measured by a free-falling clock is an example of an affine parameter, as is the distance measured by the tick marks on a free-falling ruler. An affine parameter can only be defined along a straight world-line, not an arbitrary curve. The affine measurement of 1-volume violates V1, because it only allows us to compare distances that lie on ℓ or parallel to it. On the other hand, it has the advantage over metric measurement that it allows us to measure lengths along lightlike lines.

Figure d shows how to define an affine measure of 2-volume, and a similar method works for 3-volume.

Linearity

Suppose that a parallelogram is formed with vectors **a** and **b** as two of its sides. If we double **a**, then the area doubles as well,

$$\text{area}(2\mathbf{a}, \mathbf{b}) = 2 \text{area}(\mathbf{a}, \mathbf{b}).$$

In general, if we scale either of the vectors by a factor c , the area scales by the same factor, provided that we set some rule for handling signs — an issue that we'll postpone until section 7.6.2. Something similar happens when we add two vectors, e.g.,

$$\text{area}(\mathbf{a}, \mathbf{b} + \mathbf{c}) = \text{area}(\mathbf{a}, \mathbf{b}) + \text{area}(\mathbf{a}, \mathbf{c}),$$

again postponing issues with signs. We refer to these properties as *linearity* of the affine 2-volume. Any sensible measure of m -volume should have similar linearity properties.

Change of basis

Because we have not made use of the metric so far, all of our measures of area have been relative rather than absolute. As shown in figure f, they depend on what parallelogram we choose as our unit of area. The unit cell in f/2 is smaller than the one in f/1, for two reasons: the vectors defining the edges are shorter, and the angle between them is smaller. Words like “shorter” and “angle” show us resorting to metric measurement, but we could also perform the comparison without using the metric, simply by using parallelogram 1 to measure parallelogram 2, or 2 to measure 1. If we think of such a pair of vectors as basis vectors for the plane, then switching our choice of unit parallelogram is equivalent to a change of basis. Areas change in proportion to the determinant of the matrix specifying the change of basis.

Suppose that $\mathbf{a}' = \mathbf{a}/2$, and $\mathbf{b}' = \mathbf{b}/2$. The change of basis from the unprimed pair to the primed pair is given by the matrix

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

which has determinant 4. Scaling down both basis vectors by a factor of 2 has caused a reduction by a factor of 4 in the area of the unit parallelogram. If we use the primed parallelogram to measure other areas, then all the areas will come out bigger by a factor of 4.

Rotations and Lorentz boosts are changes of basis. They have determinants equal to 1, i.e., they preserve spacetime volume.

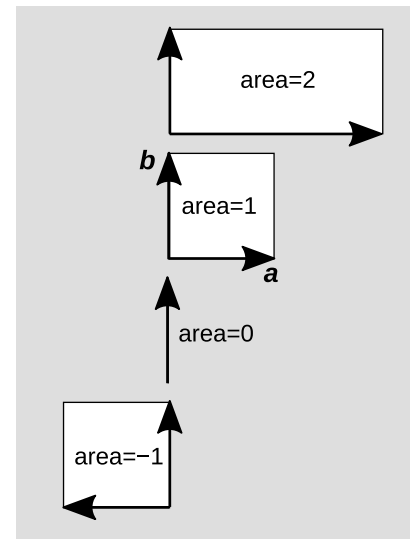
7.6.2 Orientation

As shown in figure g, linearity of area requires that some areas be assigned negative values. If we compare the areas +1 and -1 , we see that the only difference is one of orientation, or handedness. In the case to which we have arbitrarily assigned area +1, vector \mathbf{b} lies counterclockwise from vector \mathbf{a} , but when \mathbf{a} is flipped, the relative orientation becomes clockwise.

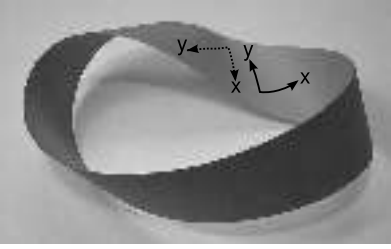
If you've had the usual freshman physics background, then you've seen this issue dealt with in a particular way, which is that we assume a third dimension to exist, and define the area to be the vector cross product $\mathbf{a} \times \mathbf{b}$, which is perpendicular to the plane inhabited by \mathbf{a} and \mathbf{b} . The trouble with this approach is that it only works in three dimensions. In four dimensions, suppose that \mathbf{a} lies along the x axis, and \mathbf{b} along the t axis. Then if we were to define $\mathbf{a} \times \mathbf{b}$, it should be in a direction perpendicular to both of these, but we have more than one such direction. We could pick anything in the y - z plane.

To get started on this issue in m dimensions, where m does not necessarily equal 3, we can consider the m -volume of the m -dimensional parallelepiped spanned by m vectors. For example, suppose that in 4-dimensional spacetime we pick our m vectors to be the unit vectors lying along the four axes of the Minkowski coordinates, $\hat{\mathbf{t}}$, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$. From experience with the vector cross product, which is anticommutative, we expect that the sign of the result will depend on the order of the vectors, so let's take them in that order. Clearly there are only two reasonable values we could imagine for this volume: +1 or -1 . The choice is arbitrary, so we make an arbitrary choice. Let's say that it's +1 for this order. This amounts to choosing an *orientation* for spacetime.

A hidden and nontrivial assumption was that once we made this choice at one point in spacetime, it could be carried over to other regions of spacetime in a consistent way. This need not be the case,



g / Linearity of area requires that some areas be assigned negative values.



h / A Möbius strip is not an orientable surface.



i / Tullio Levi-Civita (1873-1941) worked on models of number systems possessing infinitesimals and on differential geometry. He invented the tensor notation, which Einstein learned from his textbook. He was appointed to prestigious endowed chairs at Padua and the University of Rome, but was fired in 1938 because he was a Jew and an anti-fascist.

as suggested in figure h. However, our topic at the moment is special relativity, and as discussed briefly on p. 48, it is usually assumed in special relativity that spacetime is topologically trivial, so that this issue arises only in general relativity, and only in spacetimes that probably are not realistic models of our universe.

Since 4-volume is invariant under rotations and Lorentz transformations, our choice of an orientation suffices to fix a definition of 4-volume that is a Lorentz invariant. If vectors **a**, **b**, **c**, and **d** span a 4-parallelepiped, then the linearity of volume is expressed by saying that there is a set of coefficients ϵ_{ijkl} such that

$$V = \epsilon_{ijkl} a^i b^j c^k d^l.$$

Notating it this way suggests that we interpret it as abstract index notation, in which case the lack of any indices on V means that it is not just a Lorentz invariant but also a scalar.³

Halfling coordinates

Example 5

Let (t, x, y, z) be Minkowski coordinates, and let $(t', x', y', z') = (2t, 2x, 2y, 2z)$. Let's consider how each of the factors in our volume equation is affected as we do this change of coordinates.

$$\underbrace{V}_{\text{no change}} = \underbrace{\epsilon_{\kappa\lambda\mu\nu}}_{\times 1/16} \underbrace{a^\kappa}_{\times 2} \underbrace{b^\lambda}_{\times 2} \underbrace{c^\mu}_{\times 2} \underbrace{d^\nu}_{\times 2}$$

Since our convention is that V is a scalar, it doesn't change under a change of coordinates. This forces us to say that the components of ϵ change by a factor of 1/16 in this example.

The result of example 5 tells us that under our convention that volume is a scalar, the components of ϵ must change when we change coordinates. One could argue that it would be more logical to think of the transformation in this example as a change of units, in which case the value of V would be different in the new units; this is a possible alternative convention, but it would have the disadvantage of making it impossible to read off the transformation properties of an object from the number and position of its indices. Under our convention, we can read off the transformation properties in this way. Although section 7.4 only presented these properties in the case of tensors of rank 0 and 1, deferring the general description of higher-rank tensors to sec. 9.2.4, p. 180, ϵ 's transformation properties are, as implied by its four subscripts, those of a tensor of rank 4. Different authors use different conventions regarding the definition of ϵ , which was originally described by the mathematician Levi-Civita. Since by our convention ϵ is a tensor, we refer to it as the Levi-Civita tensor. In other conventions, where ϵ is not a tensor, it may be referred to as the Levi-Civita symbol. Since the notation is not standardized, I will occasionally put a reminder next to important equations containing ϵ stating that this is the tensorial ϵ .

³For the distinction, see p. 127.

The Levi-Civita tensor has lots and lots of indices. Scary! Imagine the complexity of this beast. (Sob.) We have four choices for the first index, four for the second, and so on, so that the total number of components is 256. Wait, don't reach for the kleenex. The following example shows that this complexity is illusory.

Volume in Minkowski coordinates *Example 6*
 We've set up our definitions so that for the parallelepiped $\hat{\mathbf{t}}, \hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$, we have $V = +1$. Therefore

$$\epsilon_{txyz} = +1$$

by definition, and because 4-volume is Lorentz invariant, this holds for *any* set of Minkowski coordinates.

If we interchange x and y to make the list $\hat{\mathbf{t}}, \hat{\mathbf{y}}, \hat{\mathbf{x}}, \hat{\mathbf{z}}$, then as in figure g, the volume becomes -1 , so

$$\epsilon_{tyxz} = -1.$$

Suppose we take the edges of our parallelepiped to be $\hat{\mathbf{t}}, \hat{\mathbf{x}}, \hat{\mathbf{x}}, \hat{\mathbf{z}}$, with y omitted and x duplicated. These four vectors are not linearly independent, so our parallelepiped is degenerate and has zero volume.

$$\epsilon_{txxz} = 0.$$

From these examples, we see that once any element of ϵ has been fixed, all of the others can be determined as well. The rule is that interchanging any two indices flips the sign, and any repeated index makes the result zero.

Example 6 shows that the fancy symbol ϵ_{ijkl} , which looks like a secret Mayan hieroglyph invoking 256 different numbers, actually encodes only *one* number's worth of information; every component of the tensor either equals this number, or minus this number, or zero. Suppose we're working in some set of coordinates, which may not be Minkowski, and we want to find this number. A complicated way to find it would be to use the tensor transformation law for a rank-4 tensor (sec. 9.2.4, p. 180). A much simpler way is to make use of the determinant of the metric, discussed in example 1 on p. 127. For a list of coordinates $ijkl$ that are sorted out in the order that we define to be a positive orientation, the result is simply $\epsilon_{ijkl} = \sqrt{|\det g|}$. The absolute value sign is needed because a relativistic metric has a negative determinant.

Cartesian coordinates and their halfling versions *Example 7*
 Consider Euclidean coordinates in the plane, so that the metric is a 2×2 matrix, and ϵ_{ij} has only two indices. In standard Cartesian coordinates, the metric is $g = \text{diag}(1, 1)$, which has $\det g = 1$. The Levi-Civita tensor therefore has $\epsilon_{xy} = +1$, and its other three components are uniquely determined from this one by the rules

discussed in example 6. (We could have flipped all the signs if we had wanted to choose the opposite orientation for the plane.) In matrix form, these rules result in

$$\epsilon = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Now transform to coordinates $(x', y') = (2x, 2y)$. In these coordinates, the metric is $g' = \text{diag}(1/4, 1/4)$, with $\det g = 1/16$, so that $\epsilon_{x'y'} = 1/4$, or in matrix form,

$$\epsilon' = \begin{pmatrix} 0 & 1/4 \\ -1/4 & 0 \end{pmatrix}.$$

Polar coordinates

Example 8

In polar coordinates (r, θ) , the metric is $g = \text{diag}(1, r^2)$ (problem 1, p. 160), which has determinant r^2 . The Levi-Civita tensor is

$$\epsilon = \begin{pmatrix} 0 & r \\ -r & 0 \end{pmatrix}$$

(taking the same orientation as in example 7).

Area of a circle

Example 9

Let's find the area of the unit circle. Its (signed) area is

$$A = \int 2\text{-volume}(\mathbf{dr}, \mathbf{d\theta}),$$

where the order of \mathbf{dr} and $\mathbf{d\theta}$ is chosen so that, with the orientation we've been using for the plane, the result will come out positive. Using the definition of the Levi-Civita tensor, we have

$$\begin{aligned} A &= \int \epsilon_{r\theta} dx^r dx^\theta \\ &= \int_{r=0}^1 \int_{\theta=0}^{2\pi} r dr d\theta \quad [\text{example 8}] \\ &= \pi \end{aligned}$$

7.6.3 The 3-volume covector

Consider the volume of a three-dimensional subspace of four-dimensional spacetime. Linearity leads to an especially simple characterization of the 3-volume. Let a 3-volume be defined by the parallelepiped spanned by vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} . If we threw in a fourth vector \mathbf{d} , we would have a 4-volume, and 4-volume is a scalar. This 4-volume would depend in a linear way on all four vectors, and in particular it would depend linearly on \mathbf{d} . But this means we have a scalar that is a linear function of a vector, and such a function

is exactly what we mean by a covector. We can therefore define a volume covector \mathbf{S} according to

$$S_l d^l = 4\text{-volume}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$$

or

$$S_l = \epsilon_{ijkl} a^i b^j c^k. \quad [\text{tensorial } \epsilon]$$

The volume covector collects the information about the volume of the 3-parallelepiped, encapsulating it in a convenient form with known transformation properties. In particular, the statement and proof of Gauss's theorem in 3 + 1 dimensions are greatly simplified by the use of this tool (p. 190). The 3-volume covector, unlike the affine 3-volume, is defined in an absolute sense rather than in relation to some parallelepiped arbitrarily chosen as a standard. Both the covector and the affine volume fail to satisfy the ratio-comparison property V1 on p. 151, since we can't compare volumes unless they lie in parallel 3-planes.

We've been visualizing covectors in n dimensions as stacks of $(n - 1)$ -dimensional planes (figure a/2, p. 129; figure d/2, p. 136). The volume three-vector should therefore be visualized as a stack of 3-planes in a four-dimensional space. Since most of us can't visualize things very well in four dimensions, figure j omits one of the dimensions, so that the 3-surfaces appear as two-dimensional planes. The small hand j/1 has a certain 3-volume, and the covector that measures it is represented by the stack of 3-planes parallel to it, j/2. The bigger hand j/3 has twice the 3-volume, and its covector is represented by a stack of planes with half the spacing.

If we step down from four dimensions to three, then the volume covector formed by vectors \mathbf{u} and \mathbf{v} becomes the vector cross product $\mathbf{S} = \mathbf{u} \times \mathbf{v}$, i.e., $S_k = \epsilon_{ijk} u^i v^j$.

A vector cross product

Example 10

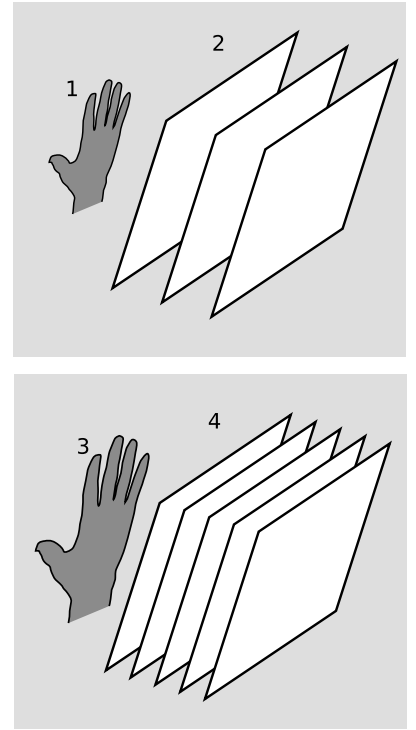
Consider Euclidean 3-space in Cartesian coordinates. We know from freshman physics that

$$\hat{\mathbf{z}} = \hat{\mathbf{x}} \times \hat{\mathbf{y}}.$$

Reexpressing this in the notation above, we have $u^x = 1$, $v^y = 1$, and zero for all the other components of \mathbf{u} and \mathbf{v} . Since the Levi-Civita tensor vanishes if we have any duplicated indices, its only nonvanishing component that can be relevant here is $\epsilon_{xyz} = 1$. (Here we assume the standard right-handed orientation for Cartesian coordinates, and we make use of the fact that $g = \text{diag}(1, 1, 1)$, so that $\det g = 1$.) The result is

$$S_z = \epsilon_{xyz} u^x v^y = 1,$$

as expected. (It doesn't matter here whether we talk about S_z or S^z , because with this metric, raising and lowering indices doesn't change the components of a vector.)



j / Interpretation of the 3-volume covector.

Classification of 3-surfaces

A useful application of the 3-volume covector is in classifying 3-surfaces by how they relate to the light cone. If I nail together three sticks, all at right angles to one another, then I can consider them as a set of basis vectors spanning a three-dimensional space of events. This three-space is flat, so we can call it a hyperplane — or just a plane if, as throughout this section, there is no danger of forgetting that it has three dimensions rather than two. All of the events in this plane are simultaneous in my frame of reference. None of these facts depends on the use of right angles; we just need to make sure that the sticks don't all lie in the same plane.

The business of a physicist is ultimately to make predictions. That is, if given a set of initial conditions, we can say how our system will evolve through time. These initial conditions are in principle measured throughout all of space, and a plane of simultaneity would be a natural choice for the set of points at which to take the measurements. A surface used for this purpose is called a Cauchy surface.

If a plane is a surface of simultaneity according to some observer, then we call it *spacelike*. Any particle's world-line must intersect such a plane exactly once, and this is why it works as a Cauchy surface: we are guaranteed to detect the particle, so that we can account for its effect on the evolution of the cosmos. We could take a spacelike plane and reorient it. For a small enough change in the orientation (that is, a change that could be described by small enough changes in the basis vectors), it will remain spacelike.

When a plane is not spacelike, and remains so under any sufficiently small change in orientation, we call it *timelike*. In Minkowski coordinates, an example would be the t - x - y plane. A given particle's world-line might never cross such a surface, and therefore a timelike plane cannot be used as a Cauchy surface.

A plane that is neither spacelike nor timelike is called *lightlike*. An example is the surface defined by the equation $x = t$ in Minkowski coordinates.

The above classification can be stated very succinctly by using the 3-volume covector defined in section 7.6.1. A plane is spacelike, lightlike, or timelike, respectively, if the regions it contains are described by 3-volume covectors that are timelike, lightlike, or spacelike. A surface that is smooth but not necessarily flat can be described locally according to these categories by considering its tangent plane. For example, a light cone is lightlike at each of its points, and since it is lightlike everywhere, we call it a lightlike surface. The event horizon of a black hole is also a lightlike surface. Any spacelike surface, whether curved or flat, can be used as a Cauchy surface.

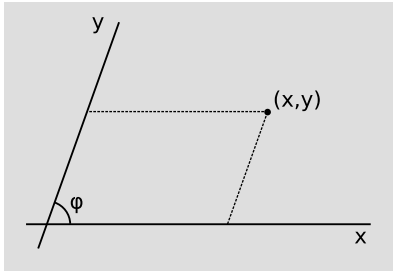
Lightlike surfaces have some funny properties. Using birdtracks notation, suppose that we form such a surface as the space spanned by the three basis vectors $\rightarrow{\mathbf{a}}$, $\rightarrow{\mathbf{b}}$, and $\rightarrow{\mathbf{c}}$, and let $\mathbf{S}\rightarrow$ be the corresponding 3-volume covector. The surface is lightlike, so

$$\mathbf{S}\rightarrow\mathbf{S} = 0. \quad (3)$$

Because $\mathbf{S}\rightarrow$ is defined as the function giving the 4-volume of a parallelepiped spanned by the bases with a fourth vector $\rightarrow{\mathbf{d}}$, and because this volume vanishes when $\rightarrow{\mathbf{d}}$ is tangent to the surface (property V2, p. 151), we have,

$$\mathbf{S}\rightarrow\mathbf{a} = \mathbf{S}\rightarrow\mathbf{b} = \mathbf{S}\rightarrow\mathbf{c} = 0. \quad (4)$$

So in this sense $\mathbf{S}\rightarrow$ is perpendicular to the surface. In Euclidean space we are used to describing the orientation of a surface in terms of the unit normal vector, and this is very nearly what $\mathbf{S}\rightarrow$ is, except that it's a covector rather than a vector, and it also can't be made to have unit length, since its magnitude is zero. We could fix the first of these two problems by constructing the vector $\rightarrow{\mathbf{S}}$ that is dual to $\mathbf{S}\rightarrow$, but this has a disconcerting effect. Combining (3) with the definition of $\mathbf{S}\rightarrow$, we find that $\rightarrow{\mathbf{S}}$ spans a vanishing 4-volume with the basis vectors, and therefore by V2 we find that $\rightarrow{\mathbf{S}}$ is tangent to the surface. Thus in some sense we have a vector that is both parallel to and tangent to a surface — which avoids being absurd because we are really referring to two *different* objects, the covector $\mathbf{S}\rightarrow$ and the vector $\rightarrow{\mathbf{S}}$.



Problem 2.

Problems

1 Example 3 on p. 149 discussed polar coordinates in the Euclidean plane. Use the technique demonstrated in section 7.3 to find the metric in these coordinates.

2 Oblique Cartesian coordinates are like normal Cartesian coordinates in the plane, but their axes are at an angle $\varphi \neq \pi/2$ to one another. Show that the metric in these coordinates is

$$ds^2 = dx^2 + dy^2 + 2 \cos \varphi \, dx \, dy.$$

3 Let a 3-plane U be defined in Minkowski coordinates by the equation $x = t$. Is this plane spacelike, timelike, or lightlike? Find a covector $\mathbf{S} \rightarrow$ that is normal to U in the sense described on p. 158, describing it in terms of its components. Compute the vector $\rightarrow \mathbf{S}$, also in component form. Verify that $\mathbf{S} \rightarrow \mathbf{S} = 0$. Show that $\rightarrow \mathbf{S}$ is tangent to M .

4 For the oblique Cartesian coordinates defined in problem 2, use the determinant of the metric to show that the Levi-Civita tensor is

$$\epsilon = \begin{pmatrix} 0 & \sin \varphi \\ -\sin \varphi & 0 \end{pmatrix}.$$

5 Use the technique demonstrated in example 9, p. 156, to find the volume of the unit sphere.

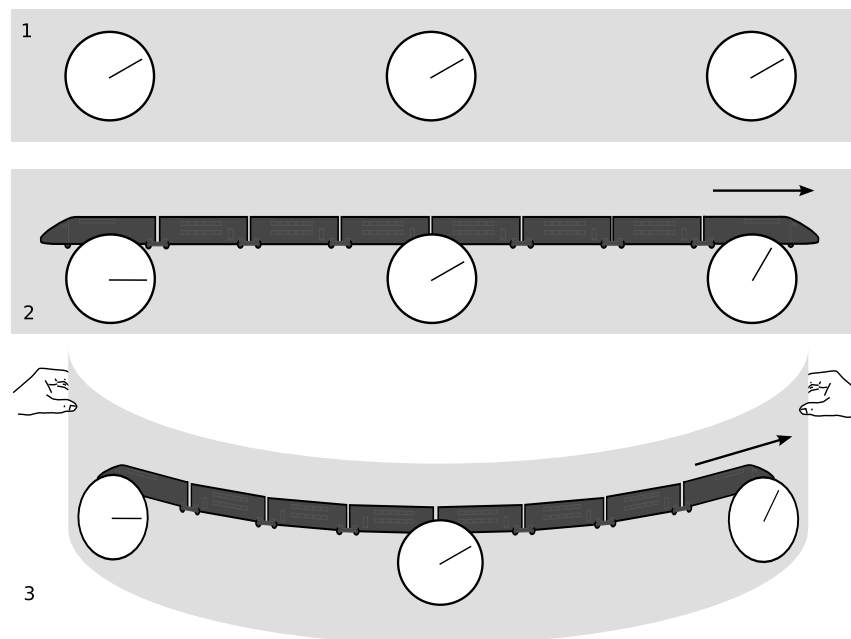
Chapter 8

Rotation (optional)

8.1 Rotating frames of reference

8.1.1 No clock synchronization

Panels 1 and 2 of figure a recapitulate the result of example 16 on p. 34. The set of three clocks fixed to the earth in a/1 have been synchronized by Einstein synchronization (example 4, p. 18), i.e., by exchanging flashes of light. The three clocks aboard the moving train, a/2, have been synchronized in the same way, and the events that were simultaneous according to frame 1 are not simultaneous in frame 2. There is a systematic shift in the times, which is represented by the term $t' = \dots - v\gamma x$ in the Lorentz transformation (eq. (1), p. 31).



a / Clocks can't be synchronized in a rotating frame of reference.

Now suppose we take the diagram of the train and wrap it around, a/3. If we go on and close the loop, making the chain into a circle like a chain necklace, we have a problem. The trend in the clock times can continue until it wraps back around to the beginning, but then there will be a discrepancy.

We conclude that clocks can't be synchronized in a rotating frame of reference. Such a frame does not admit a universal time coordinate because Einstein synchronization isn't transitive: synchronizing clock A with clock B, and B with C, does not imply that A is synchronized with C. This nontransitivity is one way of *defining* what we mean by rotation. That is, if the operational definition of an inertial frame given in section 5.1, p. 117, shows that our frame is noninertial, and we want to know more about why it's noninertial, testing for this nontransitivity is a way of finding out whether it's because of rotation.

8.1.2 Rotation is locally detectable

The people aboard the circular train know that their attempts at synchronization fail, so they can tell, without reference to anything external, that they're going in a circle. (Cf. example 1, p. 118.)

Although this is a book on special, not general, relativity, it's interesting to note the following possibility. Suppose that we verify, by local experiments, that we have a good, nonrotating, inertial frame of reference. It is then imaginable that if we view distant galaxies from this frame, we will see them rotate at some angular frequency Ω about some axis on the celestial sphere. If this is observed, then we must infer that it is the universe as a whole — not our laboratory! — that is rotating. Such an effect has been searched for, and, for example, an upper limit $\Omega \lesssim 10^{-7}$ radian/year was inferred by Clemence.¹ General-relativistic models of such rotating cosmologies have a preferred vector constituting the direction of the axis about which matter rotates, but there is no global center of rotation. Current upper limits on Ω are good enough to rule out any significant effect on cosmological expansion due to centrifugal forces.

8.1.3 The Sagnac effect

Although the train scenario is obviously unrealistic, the time shift is far from hypothetical. This type of effect, called the Sagnac effect, was first observed by M. Georges Sagnac in 1913, and it relates to the principle of the ring laser gyroscope (example 2, p. 18), used in passenger jets. (The name is French, and is pronounced “sah-NYAHK.”) To find the Sagnac effect quantitatively, we note that in the circular train example (ignoring signs) the relevant term in the Lorentz transformation, $v\gamma x$, would accumulate, after one complete circuit of Einstein synchronization, a discrepancy δ equal to the circumference of the circle multiplied by $v\gamma$. If the circle's radius is r and the angular velocity ω , we have $\Delta t = 2\pi\gamma r^2\omega$. This can be rewritten in terms of the circle's area A as $\Delta t = 2A\omega$, or, reinserting factors of c to accommodate SI units, $\Delta t = 2A\omega/c^2$. The proportionality to the enclosed area is not an accident; the product vx has the form of the integrand $\mathbf{F} \cdot d\mathbf{s}$ occurring in Stokes' theorem.

¹“Astronomical time,” Rev. Mod. Phys. 29 (1957) 2.

A clock at the equator of the earth rotates at a frequency ω of 2π radians per sidereal day, suffering a Sagnac effect of 210 ns per day. The traveling atomic clocks in the Hafele-Keating experiment (p. 15) went around the world in both directions, and were compared with a third set of clocks that stayed in Washington, DC. Since the time required to fly around the earth was also on the order of one day, the differences in the values of ω for the three sets of clocks were on the same order of magnitude as the ω of the earth, and we therefore expect cumulative differential Sagnac effects that are also on the order of a hundred nanoseconds. These effects exist only in the rotating frame of the earth, but the things being measured are proper times, and proper time is a scalar, so the experimental results are independent of what frame of reference is used for calculating them. Since the airline pilots provided Hafele and Keating with navigational data referred to the rotating earth, they analyzed their results in the rotating frame, in which there was a Sagnac effect. They could equally well have transformed their data into the frame of the stars, in which case the same result would have been predicted, but it would have been described as arising from kinematic time dilation.

The ring laser gyroscope in the photo in example 2 on p. 18 looks like it has an area on the order of 10^2 cm^2 and uses red light. For use in navigation, one wants to be able to detect a change in course of, say, one degree in our hour, or $\omega \sim 5 \times 10^{-6} \text{ radian/s}$. The result is a time shift $\Delta t \sim 10^{-24} \text{ s}$, which for red light is a phase shift of only $\Delta\phi = 4\pi A\omega/c\lambda \sim 3 \times 10^{-9} \text{ radian}$. In the original nineteenth-century experiments, this phase shift would have had to be measured by producing interference between the two beams and measuring the change in intensity resulting from this change in phase. Our estimate of ϕ shows that this is impractical for a portable instrument. In a modern ring laser gyroscope, an active laser medium is inserted in the loop, and the result is that the loop resonates at a frequency that is shifted from the laser's natural frequency by $\Delta f \sim \Delta\phi c/L$, where L is the circumference. The result is a frequency shift of a few Hz, which is easily measurable. An alternative technique, used in the fiber optic gyroscope, is to wrap N turns of optical fiber around the circumference, effectively changing A to NA .

8.1.4 A rotating coordinate system

The GPS system is a practical example of a case where we naturally want to employ a rotating coordinate system. Hikers and sailors, after all, want to know where they are relative to the earth's rotating surface. Since locations need to be determined to within meters, the timing of signals needs to be done to a precision of

something like $(1 \text{ m})/c$, which is a few nanoseconds. This is why the GPS satellites have atomic clocks aboard, and timing to this precision clearly requires that relativistic effects be taken into account. We therefore need not a rotating Newtonian coordinate system but a rotating relativistic one. Let's start with the *nonrotating* frame, and define coordinates (t, r, θ, z) , with the spatial part (r, θ, z) being ordinary cylindrical coordinates. For simplicity, we'll neglect the z coordinate in what follows. Extending the result of problem 1 on p. 160 from $2 + 0$ dimensions to $2 + 1$, we have the metric

$$ds^2 = dt^2 - dr^2 - r^2 d\theta^2. \quad (1)$$

The results of section 8.1.1 show that we do not expect to be able to define a completely satisfactory time coordinate in the rotating frame, so let's start with the minimal change $(t, r, \theta) \rightarrow (t, r, \theta')$, where $\theta' = \theta - \omega t$. This is at least enough to make world-lines of constant θ' be ones that revolve around the origin at the appropriate frequency. Substituting $d\theta = d\theta' + \omega dt$, we find

$$ds^2 = (1 - \omega^2 r^2) dt^2 - dr^2 - r^2 d\theta'^2 - 2\omega r^2 d\theta' dt. \quad (2)$$

Recognizing ωr as the velocity of one frame relative to another, and $(1 - \omega^2 r^2)^{-1/2}$ as γ , we see that we do have a relativistic time dilation effect in the dt^2 term. But the dr^2 and $d\theta'^2$ terms look the same as in equation (1). Why don't we see any Lorentz contraction of the length scale in the azimuthal direction?

The answer is that coordinates in relativity are arbitrary, and just because we can write down a certain set of coordinates, that doesn't mean they have any special physical interpretation. The coordinates (t, r, θ') do not correspond physically to the quantities that a rotating observer R would measure with clocks and meter-sticks. If R uses a ruler to measure a short arc along the circumference of the circle $r = r_0$, the distance is a distance being measured between events in spacetime that are simultaneous in the rest frame of the ruler, and these do not occur at the time value of the time coordinate t . In the Lorentz transformation, for linear motion, it is the $-v\gamma x$ term applied to the times that fixes this problems and makes t' properly represent simultaneity in the new frame. In our rotational version, we could try to do something similar by defining a time coordinate $t' = t + f\theta'$, where f is a function of r that is engineered so that the $d\theta' dt$ cross term in the metric would go away. This can be done (the function f that works turns out to be $\omega r^2/(1 - \omega^2 r^2)$), but the problem is that the t' coordinate is not single-valued, in the sense that (t, r, θ) and $(t, r, \theta + 2\pi)$ would not produce the same t' . This is inevitable, as we've seen in section 8.1.1, so we can't improve on the coordinates (t, r, θ') and the metric (2).

The coordinates (t, r, θ') , with the metric (2) are the ones used in the GPS system, and in that context are called Earth-Centered

Inertial (ECI) coordinates. (Another name is Born coordinates.) Their time coordinate is not the time measured by a clock in the rotating frame but is simply the time coordinate of the nonrotating frame of reference tied to the earth's center. Conceptually, we can imagine this time coordinate as one that is established by sending out an electromagnetic “tick-tock” signal from the earth's center, with each satellite correcting the phase of the signal based on the propagation time inferred from its own r . In reality, this is accomplished by communication with a master control station in Colorado Springs, which communicates with the satellites via relays at Kwajalein, Ascension Island, Diego Garcia, and Cape Canaveral.

8.2 Angular momentum

Nonrelativistically, the angular momentum of a particle with momentum \mathbf{p} , at a position \mathbf{r} relative to some arbitrarily fixed point, is $\mathbf{L} = \mathbf{r} \times \mathbf{p}$. When we generalize this equation to relativity, we run into a number of issues. Issues due to special relativity:

1. The vector cross product only makes sense in three dimensions, so it is not well defined in special relativity (sec. 7.6.2, p. 153).
2. Assuming we get around issue number 1, how do we know that this quantity is conserved?

And from general relativity:

3. In general relativity, only infinitesimally small spatial or space-time displacements $d\mathbf{r}$ can be treated as vectors. Larger ones cannot. This is because spacetime can be curved, and vectors can't be used to define displacements on a curved space (e.g., the surface of the earth).
4. If space has a nontrivial topology, then we may not be able to define an orientation (sec. 7.6.2, p. 153).

For points 3 and 4, we refer to Hawking and Ellis, p. 62. Number 2 is addressed in sec. 9.3.5, p. 193. For number 2 we will need the stress-energy tensor, which will be described in ch. 9. Lest you feel totally cheated, we will resolve issue number 1 in section 8.2.2, p. 167, but before we do that, let's consider an interesting example that can be handled with simpler math.

8.2.1 The relativistic Bohr model

If we want to see an interesting real-world example of relativistic angular momentum, we need something that rotates at relativistic velocities. At large scales we have astrophysical examples such as neutron stars and the accretion disks of black holes, but these involve gravity and would therefore require general relativity. At microscopic scales we have systems such as hadrons, nuclei, atoms, and

molecules. These are quantum-mechanical, and relativistic quantum mechanics is a difficult topic that is beyond the scope of this book, but we can sidestep that issue by using the Bohr model of the atom. In the Bohr model of hydrogen, we assume that the electron has a circular orbit governed by Newton's laws, does not radiate, and has its angular momentum quantized in units of \hbar . Let's generalize the Bohr model by applying relativity.

It will be convenient to define the constant $\alpha = ke^2/\hbar$, known as the fine structure constant, where k is the Coulomb constant and e is the fundamental charge. The fine structure constant is unitless and is approximately $1/137$. It is essentially a measure of the strength of the electromagnetic interaction, and in the Bohr model it also turns out to be the velocity of the electron (in units of c) in the ground state of hydrogen. Because this velocity is small compared to 1, we expect relativistic corrections in hydrogen to be small — of relative size α^2 . But we have an interesting opportunity to get at some additional and more exciting physics if we consider a *hydrogenlike* atom, i.e., an ion with Z protons in the nucleus and only one electron. Raising Z cranks up the energy scale and therefore increases the velocity as well.

Combining the Coulomb force law with the result of example 13, p. 101, for uniform circular motion, we have $kZe^2/r = m\gamma v^2$, where the factor of γ is the relativistic correction. The electron's momentum is perpendicular to the radius vector, so we assume for the moment that (as turns out to be true), the angular momentum is given by $L = rp = mv\gamma r$, where again a relativistic correction factor of γ appears. This is quantized, so let $L = \ell\hbar$, where ℓ is an integer. Solving these equations gives

$$v = \frac{Z\alpha}{\ell}$$

$$r = \frac{\ell^2\hbar}{mZ\alpha\gamma}.$$

These differ from the nonrelativistic versions only by the factor of γ in the second equation. The electrical energy is $U = -kZe^2/r$, and the kinetic energy $K = m(\gamma - 1)$ (with $c = 1$). We will find it convenient to work with the (positive) binding energy in units of the mass of the electron. Call this quantity \mathcal{E} . After some algebra, the result is

$$\mathcal{E} = 1 - \sqrt{1 - v^2}.$$

Surprisingly, this is also the exact result given by relativistic quantum mechanics if we solve the Dirac equation for the ground state, or if we take a high-energy (nearly unbound) state with the maximum value of ℓ , as is appropriate for a semiclassical circular orbit. So we can see that even though our quantum mechanics was crude, our relativity makes some sense and gives reasonable results. For

small Z , a Taylor series approximation gives $\mathcal{E} = v^2/2 + v^4/8 + \dots$, where the fourth-order term represents the relativistic correction.

So far so good, but now what if we crank up the value of Z to make the relativistic effects strong? A very disturbing thing happens when we make $Z \gtrsim 137 \approx 1/\alpha$. In the ground state we get $v > 1$ and a complex number for \mathcal{E} . Clearly something has broken down, and our results no longer make sense. We might be inclined to dismiss this as a consequence of our crude model, but remember, our calculations happened to give the same result as the Dirac equation, which has real relativistic quantum mechanics baked in. We should take this breakdown as evidence of a real *physical* breakdown. The interpretation is as follows.

According to quantum mechanics, the vacuum isn't really a vacuum. Particle-antiparticle pairs are continually popping into existence in empty space and then reannihilating one another. Their temporary creation is a violation of the conservation of mass-energy, but only a temporary violation, and this is allowed by the time-energy form of the Heisenberg uncertainty principle, $\Delta E \Delta t \gtrsim \hbar$, as long as Δt is short. It's as though we steal some money, but the police don't catch us as long as we put it back before anyone can notice. Because these particles are only temporarily in our universe, we call them virtual particles, as opposed to real particles that have a potentially permanent existence and can be detected as blips on a Geiger counter.

But when the vacuum contains an electric field that is beyond a certain critical strength, it becomes possible to create an electron-antielectron pair, let the opposite charges separate and release energy, and pay off the energy debt without having to reannihilate the particles. This is known as “sparking the vacuum.” As of this writing, only nuclei with Z up to about 118 have been discovered, and in any case the critical Z value of $1/\alpha \approx 137$ was only a rough estimate. But by colliding heavy nuclei such as lead, one can at least temporarily form an unstable compound system with a high Z , and attempts are being made to search for the predicted effect in the laboratory.

8.2.2 The angular momentum tensor

As mentioned previously, there is no such thing as a vector cross product in four dimensions, so the nonrelativistic definition of angular momentum as $\mathbf{L} = \mathbf{r} \times \mathbf{p}$ needs to be modified to be usable in relativity.

Given a position vector r^a and a momentum vector p^b , we expect based both on units and the correspondence principle that a relativistic definition of angular momentum must be some kind of a product of the vectors. Based on the rules of index notation, we don't have much leeway here. The only products we can form are

$r^a p^b$, which is a rank-2 tensor, or $r^a p_a$, a scalar. Since nonrelativistic angular momentum is a three-vector, the correspondence principle tells us that its relativistic incarnation can't be a scalar — there simply wouldn't be enough information in a scalar to tell us the things that the nonrelativistic angular momentum vector tells us: what axis the rotation is about, and which direction the rotation is.

The tensor $r^a p^b$ also has a problem, but one that can be fixed. Suppose that in a certain frame of reference a particle of mass $m \neq 0$ is at rest at the origin. Then its position four-vector at time t is $(t, 0, 0, 0)$, and its energy-momentum vector is $(m, 0, 0, 0)$. These vectors are parallel. The tensor $r^a p^b$ is nonzero and nonconserved as time flows, but clearly we want the angular momentum of an isolated particle to be conserved. Another example would be if, at a certain moment in time, we had $r = (0, x, 0, 0)$ and $p = (E, p, 0, 0)$, with both x and p positive. This particle's motion is directly away from the origin, so its angular momentum should be zero by symmetry, but $r^a p^b$ is again nonzero.

The way to fix the problem is to force the product of the position and momentum vectors to be an antisymmetric tensor:

$$L^{ab} = r^a p^b - r^b p^a.$$

Antisymmetric means that $L^{ab} = -L^{ba}$, so that elements on opposite sides of the main diagonal are the same except for opposite signs. A quick check shows that this gives the expected zero result in both of the above examples. A component such as L^{yz} measures the amount of rotation in the y - z plane. In a nonrelativistic context, we would have described this as an x component L_x of the angular momentum three-vector, because a rotation of the y - z plane about the origin is a rotation *about* the x axis — such a rotation keeps the x axis fixed. But in four-dimensional spacetime, a rotation in the y - z plane keeps the entire t - x plane fixed, so the notion of rotation “about an axis” breaks down. (Notice the pattern: in two dimensions we rotate about a point, in three dimensions rotation is about a line, and in four dimensions we rotate about a fixed plane.) In sec. 9.3.5, p. 193, we show that L^{ab} is conserved.

If we lay the angular momentum tensor out in matrix format, it looks like this:

$$\begin{pmatrix} 0 & L^{tx} & L^{ty} & L^{tz} \\ & 0 & L^{xy} & L^{xz} \\ & & 0 & L^{yz} \\ & & & 0 \end{pmatrix}.$$

The zeroes on the main diagonal are due to the antisymmetrization in the definition. I've left blanks below the main diagonal because although those components can be nonzero, they only contain a (negated) copy of the information given by the ones above the diagonal. We can see that there are really only 6 pieces of information

in this 4×4 matrix, and we've already physically interpreted the triangular cluster of three space-space components on the bottom right.

Why do we have the row on the top, consisting of the time-space components, and what do they mean physically? A highbrow answer would be that this is something very deep having to do with the fact that, as described in section 8.3 below, rotation and linear motion are not as cleanly separated in relativity as they are in nonrelativistic physics. A more straightforward answer is that in most situations these components are actually not very interesting. Consider a cloud of particles labeled $i = 1$ through n . Then for a representative component from the top row we have the total value

$$L^{tx} = \sum t_i p_i^x - \sum x_i E_i.$$

Now suppose that we fix a certain surface of simultaneity at time t . The sum becomes

$$L^{tx} = t \sum p_i^x - \sum x_i E_i.$$

There is information here, but it's not exciting information about angular momentum, it's boring information about the position and motion of the system's center of mass. If we fix a frame of reference in which the total momentum is zero, i.e., the center of mass frame, then we have $\sum p_i^x = 0$. Let's also define the position of the center of mass as the average position weighted by mass-energy, rather than the mass-weighted average, as we would do in Newtonian mechanics. Then the sum $\sum x_i E_i$ is a constant relating to the position of the center of mass, and if we like we can make it equal zero by choosing the origin of our spatial coordinates to coincide with the center of mass.

With these choices we have a much simpler angular momentum tensor:

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ & 0 & L^{xy} & L^{xz} \\ & & 0 & L^{yz} \\ & & & 0 \end{pmatrix}.$$

If we wish, we can sprinkle some notational sugar on top of all of this using the Levi-Civita tensor ϵ described in optional section 7.6, p. 151. Let's define a new tensor *L according to

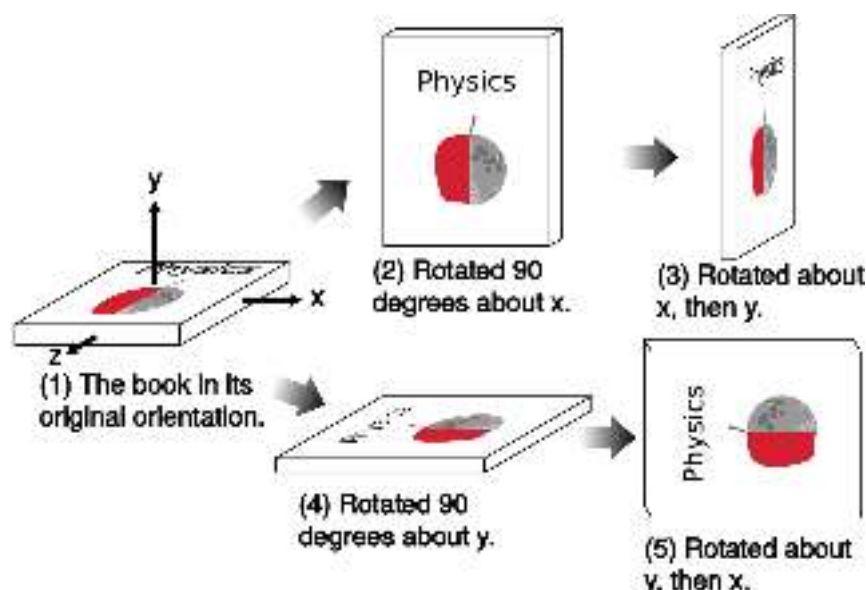
$${}^*L_{ij} = \frac{1}{2} \epsilon_{ijkl} L^{kl}.$$

Then for an observer with velocity vector o^μ , the quantity $o_\mu {}^*L^{\mu\nu}$ has the form $(0, L^{yz}, L^{zx}, L^{xy})$ (problem 3, p. 174). That is, its spatial components are exactly the quantities we would have expected for the nonrelativistic angular momentum three-vector (using the correct relativistic momentum).

8.3 Boosts and rotations

A relative of mine fell in love. She and her boyfriend bought a house in the suburbs and had a baby. They think they'll get married at some later point. An engineer by training, she says she doesn't want to get hung up on the "order of operations." For some mathematical operations, the order doesn't matter: $5 + 7$ is the same as $7 + 5$.

b / Performing the rotations in one order gives one result, 3, while reversing the order gives a different result, 5.



8.3.1 Rotations

But figure b shows that the order of operations does matter for rotations. Rotating around the x axis and then y produces a different result than y followed by x . We say that rotations are noncommutative. This is why, in Newtonian mechanics, we don't have an angular displacement vector $\Delta\theta$; vectors are supposed to be additive, and vector addition is commutative. For *small* rotations, however, the discrepancy caused by choosing one order of operations rather than the other becomes small (of order θ^2), so we *can* define an infinitesimal displacement vector $d\theta$, whose direction is given by the right-hand rule, and an angular velocity $\omega = d\theta/dt$.

As an example of how this works out for small rotations, let's take the vector

$$(0, 0, 1) \quad (3)$$

and apply the operations shown in figure b, but with rotations of only $\theta = 0.1$ radians rather than 90 degrees. Rotation by this angle about the x axis is given by the transformation $(x, y, z) \rightarrow (x, y \cos \theta - z \sin \theta, y \sin \theta + z \cos \theta)$, and applying this to the original vector gives this:

$$(0.00000, -0.09983, 0.99500) \quad (\text{after } x) \quad (4)$$

After a further rotation by the same angle, this time about the y axis, we have

$$(0.09933, -0.09983, 0.99003) \quad (\text{after } x, \text{ then } y) \quad (5)$$

Starting over from the original vector (3) and doing the operations in the opposite order gives these results:

$$(0.09983, 0.00000, 0.99500) \quad (\text{after } y) \quad (6)$$

$$(0.09983, -0.09933, 0.99003) \quad (\text{after } y, \text{ then } x) \quad (7)$$

The discrepancy between (5) and (7) is a rotation by very nearly .005 radians in the xy plane. As claimed, this is on the order of θ^2 (in fact, it's almost exactly $\theta^2/2$). A single example can never prove anything, but this is an example of the general rule that rotations along different axes don't commute, and for small angles the discrepancy is a rotation in the plane defined by the two axes, with a magnitude whose maximum size is on the order of θ^2 .

8.3.2 Boosts

Something similar happens for boosts. In $3 + 1$ dimensions, we start with the vector

$$(0, 1, 0, 0), \quad (8)$$

pointing along the x axis. A Lorentz boost with $v = 0.1$ (eq. (1), p. 31) in the x direction gives

$$(0.10050, 1.00504, 0.00000, 0.00000) \quad (\text{after } x) \quad (9)$$

and a second boost, now in the y direction, produces this:

$$(0.10101, 1.00504, 0.01010, 0.00000) \quad (\text{after } x, \text{ then } y) \quad (10)$$

Starting over from (8) and doing the boosts in the opposite order, we have

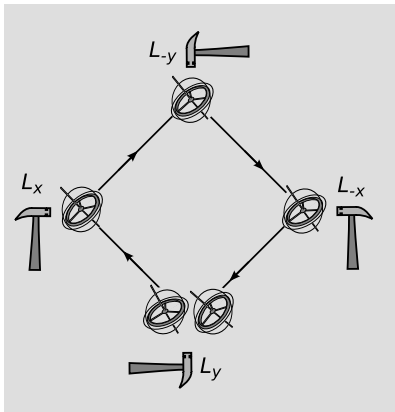
$$(0.00000, 1.00000, 0.00000, 0.00000) \quad (\text{after } y) \quad (11)$$

$$(0.10050, 1.00504, 0.00000, 0.00000) \quad (\text{after } y, \text{ then } x) \quad (12)$$

The discrepancy between (10) and (12) is a rotation in the xy plane by very nearly 0.01 radians. This is an example of a more general fact, which is that boosts along different axes don't commute, and for small angles the discrepancy is a rotation in the plane defined by the two boosts, with a magnitude whose maximum size is on the order of v^2 , in units of radians.

8.3.3 Thomas precession

Figure c shows the most important physical consequence of all this. The gyroscope is sent around the perimeter of a square, with impulses provided by hammer taps at the corners. Each impulse can be modeled as a Lorentz boost, notated, e.g., L_x for a boost



c / Nonrelativistically, the gyroscope should not rotate as long as the forces from the hammer are all transmitted to it at its center of mass.

in the x direction. The series of four operations can be written as $L_y L_x L_{-y} L_{-x}$, using the notational convention that the first operation applied is the one on the right side of the list. If boosts were commutative, we could swap the two operations in the middle of the list, giving $L_y L_{-y} L_x L_{-x}$. The L_x would undo the L_{-x} , and the L_y would undo the L_{-y} . But boosts aren't commutative, so the vector representing the orientation of the gyroscope is rotated in the xy plane. This effect is called the Thomas precession, after Llewellyn Thomas (1903-1992). Thomas precession is a purely relativistic effect, since a Newtonian gyroscope does not change its axis of rotation unless subjected to a torque; if the boosts are accomplished by forces that act at the gyroscope's center, then there is no nonrelativistic explanation for the effect.

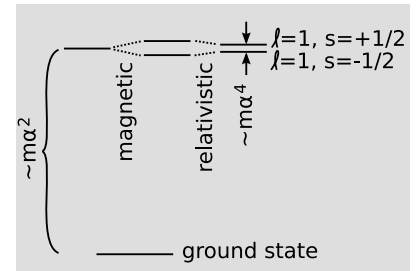
Clearly we should see the same effect if the jerky motion in figure c was replaced by uniform circular motion, and something similar should happen in any case in which a spinning object experiences an external force. In the limit of low velocities, the general expression for the angular velocity of the precession is $\Omega = \mathbf{a} \times \mathbf{v}$, and in the case of circular motion, $\Omega = (1/2)v^2\omega$, where ω is the frequency of the circular motion.

If we want to see this precession effect in real life, we should look for a system in which both v and a are large. An atom is such a system. The Bohr model, introduced in 1913, marked the first quantitatively successful, if conceptually muddled, description of the atomic energy levels of hydrogen. Continuing to take $c = 1$, the over-all scale of the energies was calculated to be proportional to $m\alpha^2$, where m is the mass of the electron, and α is the fine structure constant, defined earlier. At higher resolution, each excited energy level is found to be split into several sub-levels. The transitions among these close-lying states are in the millimeter region of the microwave spectrum. The energy scale of this fine structure is $\sim m\alpha^4$. This is down by a factor of α^2 compared to the visible-light transitions, hence the name of the constant. Uhlenbeck and Goudsmit showed in 1926 that a splitting on this order of magnitude was to be expected due to the magnetic interaction between the proton and the electron's magnetic moment, oriented along its spin. The effect they calculated, however, was too big by a factor of two.

The explanation of the mysterious factor of two had in fact been implicit in a 1916 calculation by Willem de Sitter, one of the first applications of general relativity. De Sitter treated the earth-moon system as a gyroscope, and found the precession of its axis of rotation, which was partly due to the curvature of spacetime and partly due to the type of rotation described earlier in this section. The effect on the motion of the moon was noncumulative, and was only about one meter, which was much too small to be measured at the time. In 1927, however, Thomas applied similar reasoning to the

hydrogen atom, with the electron's spin vector playing the role of gyroscope. Since the electron's spin is $\hbar/2$, the energy splitting is $\pm(\hbar/2)\Omega$, depending on whether the electron's spin is in the same direction as its orbital motion, or in the opposite direction. This is less than the atom's gross energy scale $\hbar\omega$ by a factor of $v^2/2$, which is $\sim \alpha^2$. The Thomas precession cancels out half of the magnetic effect, bringing theory in agreement with experiment.

Uhlenbeck later recalled: "...when I first heard about [the Thomas precession], it seemed unbelievable that a relativistic effect could give a factor of 2 instead of something of order v/c ... Even the cognoscenti of relativity theory (Einstein included!) were quite surprised."



d / States in hydrogen are labeled with their ℓ and s quantum numbers, representing their orbital and spin angular momenta in units of \hbar . The state with $s = +1/2$ has its spin angular momentum aligned with its orbital angular momentum, while the $s = -1/2$ state has the two angular momenta in opposite directions. The direction and order of magnitude of the splitting between the two $\ell = 1$ states is successfully explained by magnetic interactions with the proton, but the calculated effect is too big by a factor of 2. The relativistic Thomas precession cancels out half of the effect.

Problems

1 In the 1925 Michelson-Gale-Pearson experiment, the physicists measured the Sagnac effect due to the earth's rotation. They laid out a rectangle of sewer pipes with length $x = 613$ m and width $y = 339$ m, and pumped out the air. The latitude of the site in Illinois was $41^\circ 46'$, so that the effective area was equal to the projection of the rectangle into the plane perpendicular to the earth's axis. Light was provided by a sodium discharge with $\lambda = 570$ nm. The light was sent in both directions around the rectangle and interfered, effectively doubling the area. Clever techniques were required in order to calibrate the apparatus, since it was not possible to change its orientation. Calculate the number of wavelengths by which the relative phase of the two beams was expected to shift due to the Sagnac effect, and compare with the experimentally measured result of 0.230 ± 0.005 cycles.

2 The relativistic heavy ion collider RHIC collides counter-rotating beams of gold nuclei at 9 GeV/nucleon. If a gold nucleus is approximately a sphere with radius 6×10^{-15} m, find the maximum angular momentum, in units of \hbar , about the center of mass for a sideswiping collision. Answer: $\sim 10^5$.

3 Show, as claimed on p. 169, that the time-space components of the tensor $*L$ equal the angular momentum three-vector.

Chapter 9

Flux

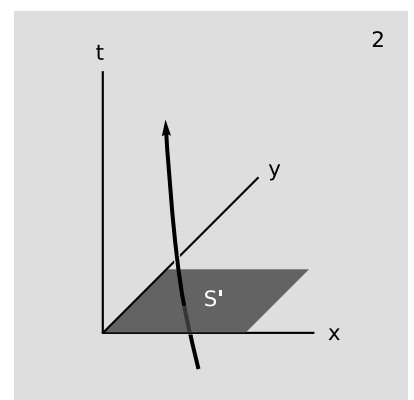
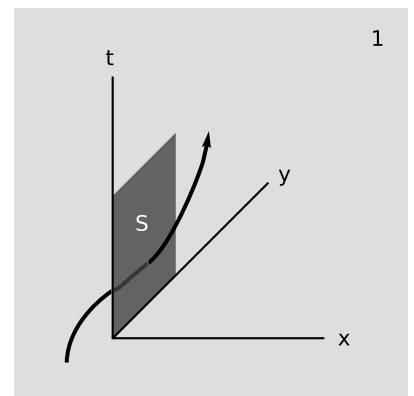
9.1 The current vector

9.1.1 Current as the flux of charged particles

The most fundamental laws of physics are conservation laws, which tell us that we can't create or destroy "stuff," where "stuff" could mean quantities such as electric charge or energy-momentum. Since charge is a Lorentz invariant, it's an easy example to start with. Because charge is invariant, we might also imagine that charge density ρ was invariant. But this is not the case, essentially because spatial (3-dimensional) volume isn't invariant; in $3 + 1$ dimensions, only *four*-dimensional volume is an invariant (problem 2, p. 51). For example, suppose we have an insulator in the shape of a cube, with charge distributed uniformly throughout it according to an observer \mathbf{o}_1 at rest relative to the cube. Then in a frame \mathbf{o}_2 moving relative to the cube, parallel to one of its axes, the cube becomes foreshortened by length contraction, and its volume is reduced by the factor $1/\gamma$. The result is that the charge density in \mathbf{o}_2 is greater by a factor of γ .

This means that knowledge of the charge density ρ in one frame is insufficient to determine the charge density in another frame. In the example of the cube, what would be sufficient would be knowledge of the vector $\mathbf{J} = \rho_0 \mathbf{v}$, where ρ_0 is the charge density in the cube's rest frame, and \mathbf{v} is the cube's velocity vector. \mathbf{J} , called the current vector, transforms as a relativistic vector because of the transformation properties of the two factors that define it. The velocity \mathbf{v} is a vector (section 3.5.1). The factor ρ_0 is an invariant, since it in turn breaks down into charge divided by rest-volume. Charge is an invariant, and all observers agree on what the volume the cube *would* have in its rest frame.

\mathbf{J} can be expressed in Minkowski coordinates as (ρ, J^x, J^y, J^z) , where ρ is the charge density and, e.g., J^x is the density of electric current in the x direction. Suppose we define the three-surface S shown in figure a/1, consisting of the set of events with coordinates $(t, 0, y, z)$ such that $0 \leq t \leq 1$, $0 \leq y \leq 1$, and $0 \leq z \leq 1$. Some charged particles have world-lines that intersect this surface, passing through it either in the positive x direction or the negative x direction (which we count as negative charge transport). S has a three-volume V . If we add up the total charge transport Δq across

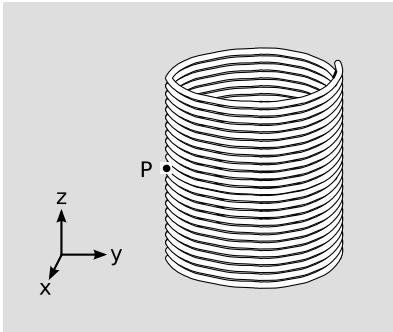


a / Charged particles with world-lines that contribute to J_x and ρ . The z dimension isn't shown, so the cubical 3-surfaces appear as squares.

this surface and divide by V , we get the average value of J^x . If we let S shrink down to smaller and smaller three-surfaces surrounding the event $(0,0,0,0)$, then we get the value of J^x at this point, $\lim_{V \rightarrow 0} \Delta q/V$. In other words, J^x measures the *flux density* of charge that passes through S . Of course this description in terms of a limit implies a large number of charges, not just one as in figure a.

You can write out the analogous definition for J^t , using a surface of simultaneity for like S' , figure a/2, and you'll see that it expresses the density of charge ρ . In this case S' represents a moment in time, and the flux through S' means that the charges are crossing the threshold from the past into the future.

Our argument that \mathbf{J} transformed like a vector was based on a case where all the charged particles had the same velocity vector, but the above description in terms of the flux of charge eliminated any discussion of velocity. It's true, but less obvious, that the \mathbf{J} described in this way also transforms as a vector, even in cases where the charged particles do not all have parallel world-lines. The current vector is the source of electric and magnetic fields. Remarkably, no macroscopic electrical measurement is capable of detecting anything more detailed about the motion of the charges than the averaged information provided by \mathbf{J} .



b / Example 1.

Boosting a solenoid

Example 1

The figure shows a solenoid, at rest, wound from copper wire. At point P, we construct a rectangular Ampèrian loop in the yz plane that has its right edge inside the solenoid and its left one outside. Ampère's law, $\int \mathbf{B} \cdot d\mathbf{s} = (4\pi k/c^2)I$, then tells us that the current density J_x causes a difference between the exterior field $B_z = 0$ and the interior field $B_z = (4\pi k/c^2)J_x \Delta y$, where Δy is the thickness of the solenoid. There are two things we can get from this result, both of them nontrivial.

First, the field depends only on the current density, not on any information about the details of the motion of the electrons in the copper. The electrons' motion is fast and highly random, but all that contributes to J_x is the slow drift velocity, typically ~ 1 cm/s, superimposed on the randomness. This is exact and not at all obvious. For example, the total *momentum* of the electrons *does* depend on the random part of their motion, because $p_x = m\gamma v_x$ has a factor of γ in it.

Second, we can use the transformation properties of the current vector to find the field of this solenoid in a frame boosted along its axis. This is the kind of situation that would naturally arise, for example, in an electric motor whose rotor contains an electromagnet. A Lorentz transformation in the z direction doesn't change the x component of a vector, nor does it change Δy , so B_z is the same in both frames. This is nontrivial both in the sense

that it would have been difficult to figure out by brute force and in the sense that fields *don't* have to be the same in different frames of reference — for example, a boost in the x or the y direction would have changed the result.

A wire

Example 2

In a solid conductor such as a copper wire, we have two types of charges, protons and electrons. The protons are at rest in the lab frame \mathbf{o} , with charge density ρ_p and current density

$$\mathbf{J}_p = (\rho_p, 0, 0, 0)$$

in Minkowski coordinates. The motion of the electrons is complicated. Some electrons are bound to a particular atom, but still move at relativistic speeds within their atoms. Others exhibit violent thermal motion that very nearly, but not quite, averages out to zero when there is a current measurable by an ammeter. For simplicity, we treat all the electrons (both the bound ones and the mobile ones) as a single density of charge ρ_e . Let the average velocity of the electrons, known as their drift velocity, be v in the x direction. Then in the frame \mathbf{o}' moving along with the drift velocity we have

$$\mathbf{J}'_e = (\rho'_e, 0, 0, 0),$$

which under a Lorentz transformation back into the lab frame becomes

$$\mathbf{J}_e = (\rho'_e \gamma, \rho'_e v \gamma, 0, 0).$$

Adding the two current vectors, we have a total current in the lab frame

$$\mathbf{J} = (\rho_p + \rho'_e \gamma, \rho'_e v \gamma, 0, 0).$$

The wire is electrically neutral in this frame, so $\rho_p + \rho'_e \gamma = 0$. Since ρ_p is a fixed property of the wire, we express ρ'_e in terms of it as ρ_p/γ . Eliminating ρ'_e gives

$$\mathbf{J} = (0, -\rho_p v, 0, 0).$$

Because the γ factors canceled, we find that the current is exactly proportional to the drift velocity. Geometrically, we have added two timelike vectors and gotten a spacelike one; this is possible because one of the timelike vectors was future-directed and the other past-directed.

9.1.2 Conservation of charge

Conservation of charge can be expressed elegantly in terms of \mathbf{J} . Charge density is the timelike component J^t . If this charge density near a certain point is, for example, increasing, then it might be because charge conservation has been violated as in figure c/1. In this example, more world-lines emerge into the future at the top of the four-cube than had entered through the bottom in the past. Some process inside the cube is creating charge. In the limit where the cube is made very small, this would be measured by a value of $\partial J^t / \partial t$ that was greater than zero.

But experiments have never detected any violation of charge conservation, so if more charge is emerging from the top (future) side of the cube than came in from the bottom (past), the more likely explanation is that the charges are not all at rest, as in c/1, but are moving, c/2, and there has been a net flow in from neighboring regions of space. We should find this reflected in the spatial components J^x , J^y and J^z . Moreover, if these spatial components were all constant, then any given region of space would have just as much current flowing into it from one side as there was flowing out the other. We therefore need to have some nonzero partial derivatives such as $\partial J^x / \partial x$. For example, figure c/2 has a positive J^x on the left and a negative J^x on the right, so $\partial J^x / \partial x < 0$. Charge conservation is expressed by the simple equation $\partial J^\lambda / \partial x^\lambda = 0$. Writing out the implied sum over λ , this says that $\partial J^t / \partial t + \partial J^x / \partial x + \partial J^y / \partial y + \partial J^z / \partial z = 0$, with an implied sum over the index λ . If you've taken vector calculus, you'll recognize the operator being applied to \mathbf{J} as a four-dimensional generalization of the divergence. This charge-conservation equation is valid regardless of the coordinate system, so it can also be rewritten in abstract index notation as

$$\frac{\partial J^a}{\partial x^a} = 0. \quad (1)$$

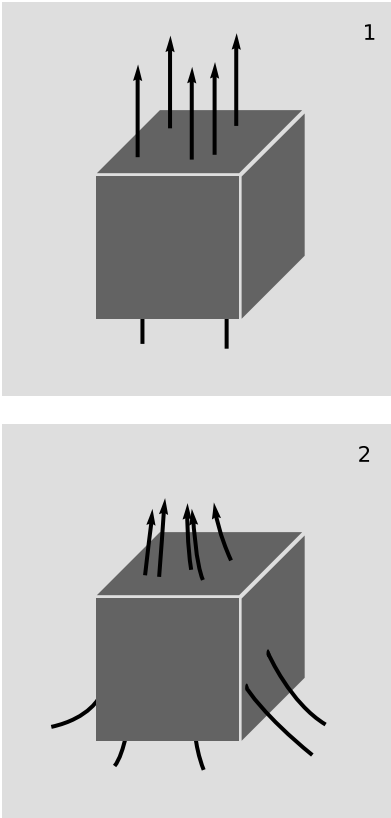
Conservation of charge in a solenoid

Example 3

In a solenoid, we have charge circulating at some drift velocity v . Ignoring the protons, and adapting the relevant expression from example 2 to the case of circular rather than linear motion, we might have for the electrons' contribution to the current something of the form

$$\mathbf{J} = p(1, -qy, qx, 0),$$

where $p = \gamma v$ and q depends on the v and on the radius of the solenoid. Conservation of charge is satisfied, because each of the four terms in the equation $\partial J^t / \partial t + \partial J^x / \partial x + \partial J^y / \partial y + \partial J^z / \partial z = 0$ vanishes individually.



c / 1. Charge is not conserved. Charges mysteriously appear at a later time without having been present before. 2. Charge is conserved. Although more world-lines come out through the top of the box than came in through the bottom, the discrepancy is accounted for by others that entered through the sides.

9.2 The stress-energy tensor

9.2.1 Conservation and flux of energy-momentum

A particle such as an electron has a charge, but it also has a mass. We can't define a relativistic mass flux because flux is defined by addition, but mass isn't additive in relativity (example 6, p. 89). Mass-energy is additive, but unlike charge it isn't an invariant. Mass-energy is part of the energy-momentum four vector $\mathbf{p} = (E, p^x, p^y, p^z)$. We then have sixteen different fluxes we can define. For example, we could replay the description in section 9.1 of the three-surface S perpendicular to the x direction, but now we would be interested in a quantity such as the z component of momentum. We then have a measure of the density of flux of p^z in the x direction, which we notate as T^{zx} . The matrix T is called the stress-energy tensor, and it is an object of central importance in relativity. (The reason for the odd name will become more clear in a moment.) In general relativity, it is the source of gravitational fields.

The stress-energy tensor is related to physical measurements as follows. Let \mathbf{o} be the future-directed, normalized velocity vector of an observer; let \mathbf{s} express a spatial direction according to this observer, i.e., it points in a direction of simultaneity and is normalized with $\mathbf{s} \cdot \mathbf{s} = -1$; and let \mathbf{S} be a three-volume covector (p. 156), directed toward the future (i.e., $o^a S_a > 0$). Then measurements by this observer come out as follows:

$$T^{ab} o_a S_b = \text{mass-energy inside the three-volume } \mathbf{S} \quad (2a)$$

$$T^{ab} s_a S_b = \text{momentum in the direction } \mathbf{s}, \text{ inside } \mathbf{S} \quad (2b)$$

The stress-energy tensor allows us to express conservation of energy-momentum as

$$\frac{\partial T^{ab}}{\partial x^a} = 0. \quad (3)$$

This *local* conservation of energy-momentum is all we get in general relativity. As discussed in section 4.3.6, p. 97, there is no such global law in curved spacetime. However, we will show in section 9.3.4 that in the special case of flat spacetime, i.e., special relativity, we do have such a global conservation law.

9.2.2 Symmetry of the stress-energy tensor

The stress-energy tensor is a symmetric matrix. For example, let's say we have some nonrelativistic particles. If we have a nonzero T^{tx} , it represents a flux of mass-energy (p^t) through a three-surface perpendicular to x . This means that mass is moving in the x direction. But if mass is moving in the x direction, then we have some x momentum p^x . Therefore we must also have a T^{xt} , since this momentum is carried by the particles, whose world-lines pass through a hypersurface of simultaneity.

9.2.3 Dust

The simplest example of a stress-energy tensor would be a cloud of particles, all at rest in a certain frame of reference, described in Minkowski coordinates:

$$T^{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

where we now use ρ to indicate the density of mass-energy, not charge as in section 9.1. This could be the stress-energy tensor of a stack of oranges at the grocery store, the atoms in a hunk of copper, or the galaxies in some small neighborhood of the universe. Relativists refer to this type of matter, in which the velocities are negligible, as “dust.” The nonvanishing component T^{tt} indicates that for a three-surface S perpendicular to the t axis, particles with mass-energy $E = P^t$ are crossing that surface from the past to the future. Conservation of energy-momentum is satisfied, since all the elements of this T are constant, so all the partial derivatives vanish.

9.2.4 Rank-2 tensors and their transformation law

Suppose we were to look at this cloud in a different frame of reference. Some or all of the timelike row $T^{t\nu}$ and timelike column $T^{\mu t}$ would fill in because of the existence of momentum, but let’s just focus for the moment on the change in the mass-energy density represented by T^{tt} . It will increase for two reasons. First, the kinetic energy of each particle is now nonzero; its mass-energy increases from m to $m\gamma$. But in addition, the volume occupied by the cloud has been reduced by $1/\gamma$ due to length contraction. We’ve picked up two factors of gamma, so the result is $\rho \rightarrow \rho\gamma^2$. This is different from the transformation behavior of a vector. When a vector is purely timelike in one frame, transformation to another frame raises its timelike component only by a factor of γ , not γ^2 . This tells us that a matrix like T transforms differently than a vector (section 7.2, p. 145). The general rule is that if we transform from coordinates x to x' , then:

$$T'^{\mu\nu} = T^{\kappa\lambda} \frac{\partial x'^{\mu}}{\partial x^{\kappa}} \frac{\partial x'^{\nu}}{\partial x^{\lambda}} \quad (4)$$

An object that transforms in this standard way is called a rank-2 tensor. The 2 is because it has two indices. Vectors and covectors have rank 1, invariants rank 0.

In section 7.3, p. 146, we developed a method of transforming the metric from one set of coordinates to another; we now see that technique as an application of the more general rule given in equation (4). Considered as a tensor, the metric is symmetric, $g_{ab} = g_{ba}$. In most of the example’s we’ve been considering, the metric tensor is diagonal, but when it has off-diagonal elements, each of these is

one half the corresponding coefficient in the expression for ds , as in the following example.

An non-diagonal metric tensor

Example 4

The answer to problem 2 on p. 160 was the metric

$$ds^2 = dx^2 + dy^2 + 2 \cos \varphi \, dx \, dy.$$

Writing this in terms of the metric tensor, we have

$$\begin{aligned} ds^2 &= g_{\mu\nu} \, dx^\mu \, dx^\nu \\ &= g_{xx} \, dx^2 + g_{xy} \, dx \, dy + g_{yx} \, dy \, dx + g_{yy} \, dy^2 \\ &= g_{xx} \, dx^2 + 2g_{xy} \, dx \, dy + g_{yy} \, dy^2. \end{aligned}$$

Therefore we have $g_{xy} = \cos \varphi$, not $g_{xy} = 2 \cos \varphi$.

Dust in a different frame

Example 5

We start with the stress-energy tensor of the cloud of particles, in the rest frame of the particles.

$$T^{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Under a boost by v in the x direction, the tensor transformation law gives

$$T^{\mu'\nu'} = \begin{pmatrix} \gamma^2 \rho & \gamma^2 v \rho & 0 & 0 \\ \gamma^2 v \rho & \gamma^2 v^2 \rho & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The over-all factor of γ^2 arises for the reasons previously described.

Parity

Example 6

The parity transformation is a change of coordinates that looks like this:

$$\begin{aligned} t' &= t \\ x' &= -x \\ y' &= -y \\ z' &= -z \end{aligned}$$

It turns right-handed screws into left-handed ones, but leaves the arrow of time unchanged. Under this transformation, the tensor transformation law tells us that some of the components of the stress-energy tensor will flip their signs, while others will stay the same:

$$\begin{pmatrix} \text{no flip} & \text{flip} & \text{flip} & \text{flip} \\ \text{flip} & \text{no flip} & \text{no flip} & \text{no flip} \\ \text{flip} & \text{no flip} & \text{no flip} & \text{no flip} \\ \text{flip} & \text{no flip} & \text{no flip} & \text{no flip} \end{pmatrix},$$

Everything here was based solely on the fact that T was a rank-2 tensor expressed in Minkowski coordinates, and therefore the same parity properties hold for other rank-2 tensors as well; cf. example 1, p. 220.

9.2.5 Pressure

The stress-energy tensor carries information about pressure. For example, T^{xx} is the flux in the x direction of x -momentum. This is simply the pressure, P , that would be exerted on a surface with its normal in the x direction. Negative pressure is tension, and this is the origin of the term “tensor,” coined by Levi-Civita (see p. 154).

Pressure as a source of gravitational fields

Example 7

Because the stress-energy tensor is the source of gravitational fields in general relativity, we can see that the gravitational field of an object should be influenced not just by its mass-energy but by its internal stresses. The very early universe was dominated by photons rather than by matter, and photons have a much higher ratio of momentum to mass-energy than matter, so the importance of the pressure components in the stress-energy tensor was much greater in that era. In the universe today, the largest pressures are those found inside atomic nuclei. Inside a heavy nucleus, the electromagnetic pressure can be as high as 10^{33} Pa! If general relativity’s description of pressure as a source of gravitational fields were wrong, then we would see anomalous effects in the gravitational forces exerted by heavy elements compared to light ones. Such effects have been searched for both in the laboratory¹ and in lunar laser ranging experiments,² with results that agreed with general relativity’s predictions.

9.2.6 A perfect fluid

The cloud in example 5 had a stress-energy tensor in its own rest frame that was isotropic, i.e., symmetric with respect to the x , y , and z directions. The tensor became anisotropic when we switched out of this frame. If a physical system has a frame in which its stress-energy tensor is isotropic, i.e., of the form

$$T^{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & P & 0 & 0 \\ 0 & 0 & P & 0 \\ 0 & 0 & 0 & P \end{pmatrix},$$

we call it a perfect fluid in equilibrium. Although it may contain moving particles, this special frame is the one in which their momenta cancel out. In other cases, the pressure need not be isotropic,

¹Kreuzer, Phys. Rev. 169 (1968) 1007. Described in section 3.7.3 of Will, “The Confrontation between General Relativity and Experiment,” relativity.livingreviews.org/Articles/lrr-2006-3/.

²Bartlett and van Buren, Phys. Rev. Lett. 57 (1986) 21, also described in Will.

and the stress exerted by the fluid need not be perpendicular to the surface on which it acts. The space-space components of T would then be the classical stress tensor, whose diagonal elements are the anisotropic pressure, and whose off-diagonal elements are the shear stress. This is the reason for calling T the stress-energy tensor.

The perfect fluid form of the stress-energy tensor is extremely important and common. For example, cosmologists find that it is a nearly perfect description of the universe on large scales.

We discussed in section ?? the ideas of converting back and forth between vectors and their corresponding covectors, and of notating this as the raising and lowering indices. We can do the same thing with the *two* indices of a rank-2 tensor, so that the stress-energy tensor can be expressed in four different ways: T^{ab} , T_{ab} , T^a_b , and T_a^b , but the symmetry of T means that there is no interesting distinction between the final two of these. In special relativity, the distinctions among the various forms are not especially fascinating. We can always cover all of spacetime with Minkowski coordinates, so that the form of the metric is simply a diagonal matrix with elements ± 1 on the diagonal. As with a rank-1 tensor, raising and lowering indices on a rank-2 tensor just flips some components and leaves others alone. The methods for raising and lowering don't need to be deduced or memorized, since they follow uniquely from the grammar of index notation, e.g., $T^a_b = g_{bc}T^{ac}$. But there is the potential for a lot of confusion with all the signs, and in addition there is the fact that some people use a $+ - -$ signature while others use $- + +$. Since perfect fluids are so important, I'll demonstrate how all of this works out in that case.

For a perfect fluid, we can write the stress-energy tensor in the coordinate-independent form

$$T^{ab} = (\rho + P)o^a o^b - (o^c o_c)Pg^{ab},$$

where \mathbf{o} represents the velocity vector of an observer in the fluid's rest frame, and $o^c o_c = o^2 = \mathbf{o} \cdot \mathbf{o}$ equals 1 for our $+ - -$ signature or -1 for the signature $- + +$. For ease of writing, let's abbreviate the signature factor as $s = o^c o_c$.

Suppose that the metric is diagonal, but its components are varying, $g_{\alpha\beta} = \text{diag}(sA^2, -sB^2, \dots)$. The properly normalized velocity vector of an observer at (coordinate-)rest is $o^\alpha = (A^{-1}, 0, 0, 0)$. Lowering the index gives $o_\alpha = (sA, 0, 0, 0)$. The various forms of the stress-energy tensor then look like the following:

$$\begin{aligned} T_{00} &= A^2 \rho & T_{11} &= B^2 P \\ T^0_0 &= s \rho & T^1_1 &= -s P \\ T^{00} &= A^{-2} \rho & T^{11} &= B^{-2} P. \end{aligned}$$

Which of these forms is the “real” one, e.g., which form of the 00 component is the one that the observer \mathbf{o} actually measures when

she sticks a shovel in the ground, pulls out a certain volume of dirt, weighs it, and determines ρ ? The answer is that the index notation is so slick and well designed that *all* of them are equally “real,” and we don’t need to memorize which actually corresponds to measurements. When she does this measurement with the shovel, she could say that she is measuring the quantity $T^{ab}o_a o_b$. But because all of the a ’s and b ’s are paired off, this expression is a rank-0 tensor. That means that $T^{ab}o_a o_b$, $T_{ab}o^a o^b$, and $T^a_b o_a o^b$ are all the same number. If, for example, we have coordinates in which the metric is diagonal and has elements ± 1 , then in all these expressions the differing signs of the o ’s are exactly compensated for by the signs of the T ’s.

9.2.7 Two simple examples

A rope under tension

Example 8

As a real-world example in which the pressure is *not* isotropic, consider a rope that is moving inertially but under tension, i.e., equal forces at its ends cancel out so that the rope doesn’t accelerate. Tension is the same as negative pressure. If the rope lies along the x axis and its fibers are only capable of supporting tension along that axis, then the rope’s stress-energy tensor will be of the form

$$T^{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & P & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

where P is negative and equals minus the tension per unit cross-sectional area.

Conservation of energy-momentum is expressed as (eq. 3, p. 179)

$$\frac{\partial T^{ab}}{\partial x^a} = 0.$$

Converting the abstract indices to concrete ones, we have

$$\frac{\partial T^{\mu\nu}}{\partial x^\mu} = 0,$$

where there is an implied sum over μ , and the equation must hold both in the case where ν is a label for t and the one where it refers to x .

In the first case, we have

$$\frac{\partial T^{tt}}{\partial t} + \frac{\partial T^{xt}}{\partial x} = 0,$$

which is a statement of conservation of energy, energy being the timelike component of the energy-momentum. The first term is zero because ρ is constant by virtue of our assumption that the rope was uniform. The second term is zero because $T^{xt} = 0$.

Therefore conservation of energy is satisfied. This came about automatically because by writing down a time-independent expression for the stress-energy, we were dictating a static equilibrium.

When ν stands for x , we get an equation that requires the x component of momentum to be conserved,

$$\frac{\partial T^{tx}}{\partial t} + \frac{\partial T^{xx}}{\partial x} = 0.$$

This simply says

$$\frac{\partial P}{\partial x} = 0,$$

meaning that the tension in the rope is constant along its length.

A rope supporting its own weight

Example 9

A variation on example 8 is one in which the rope is hanging and supports its own weight. Although gravity is involved, we can solve this problem without general relativity, by exploiting the equivalence principle (section 5.2, p. 120). As discussed in section 5.1 on p. 117, an inertial frame in relativity is one that is free-falling. We define an inertial frame of reference \mathbf{o} , corresponding to an observer free-falling past the rope, and a noninertial frame \mathbf{o}' at rest relative to the rope.

Since the rope is hanging in static equilibrium, observer \mathbf{o}' sees a stress-energy tensor that has no time-dependence. The off-diagonal components vanish in this frame, since there is no momentum. The stress-energy tensor is

$$T^{\mu'\nu'} = \begin{pmatrix} \rho & 0 \\ 0 & P \end{pmatrix},$$

where the components involving y and z are zero and not shown, and P is negative as in example 8. We could try to apply the conservation of energy condition to this stress-energy tensor as in example 8, but that would be a mistake. As discussed in 7.5 on p. 150, rates of change can *only* be measured by taking partial derivatives with respect to the coordinates if the coordinates are Minkowski, i.e., in an inertial frame. Therefore we need to transform this stress-energy tensor into the inertial frame \mathbf{o} .

For simplicity, we restrict ourselves to the Newtonian approximation, so that the change of coordinates between the two frames is

$$\begin{aligned} t &\approx t' \\ x &\approx x' + \frac{1}{2}at'^2, \end{aligned}$$

where $a > 0$ if the free-falling observer falls in the negative x direction, i.e., positive x is up. That is, if a point on the rope at a

fixed x' is marked with a spot of paint, then free-falling observer \mathbf{o} sees the spot moving up, to larger values of x , at $t > 0$. Applying the tensor transformation law, we find

$$T^{\mu\nu} = \begin{pmatrix} \rho & \rho at \\ \rho at & P + \rho a^2 t^2 \end{pmatrix},$$

As in example 8, conservation of energy is trivially satisfied. Conservation of momentum gives

$$\frac{\partial T^{tx}}{\partial t} + \frac{\partial T^{xx}}{\partial x} = 0,$$

or

$$\rho a + \frac{\partial P}{\partial x} = 0.$$

Integrating this with respect to x , we have

$$P = -\rho ax + \text{constant}.$$

Let the cross-sectional area of the rope be A , and let $\mu = \rho A$ be the mass per unit length and $T = -PA$ the tension. We then find

$$T = \mu ax + \text{constant}.$$

Conservation of momentum requires that the tension vary along the length of the rope, just as we expect from Newton's laws: a section of the rope higher up has more weight below it to support.

9.2.8 Energy conditions

The result of example 9 could cause something scary to happen. If we walk up to a clothesline under tension and give it a quick karate chop, we will observe wave pulses propagating away from the chop in both directions, at velocities $v = \pm\sqrt{T/\mu}$. But the result of the example is that this expression increases without limit as x gets larger and larger. At some point, v will exceed the speed of light. (Of course any real rope would break long before this much tension was achieved.) Two things led to the problematic result: (1) we assumed there was no constraint on the possible stress-energy tensor in the rest frame of the rope; and (2) we used a Newtonian approximation to change from this frame to the free-falling frame. In reality, we don't know of any material so stiff that vibrations propagate in it faster than c . In fact, all ordinary materials are made of atoms, atoms are bound to each other by electromagnetic forces, and therefore no material made of atoms can transmit vibrations faster than the speed of an electromagnetic wave, c .

Based on these conditions, we therefore expect there to be certain constraints on the stress-energy tensor of any ordinary form

of matter. For example, we don't expect to find any rope whose stress-energy tensor looks like this:

$$T^{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

because here the tensile stress $+2$ is greater than the mass density 1 , which would lead to $|v| = \sqrt{2/1} > 1$. Constraints of this kind are called energy conditions. Hypothetical forms of matter that violate them are referred to as exotic matter; if they exist, they are not made of atoms. This particular example violates the an energy condition known as the dominant energy condition, which requires $\rho > 0$ and $|P| > \rho$. There are about five energy conditions that are commonly used, and a detailed discussion of them is more appropriate for a general relativity text. The common ideas that recur in many of them are: (1) that energy density is never negative in any frame of reference, and (2) that there is never a flux of energy propagating at a speed greater than c .

An energy condition that is particularly simple to express is the trace energy condition (TEC),

$$T^a_a \geq 0,$$

where we have to have one upper index and one lower index in order to obey the grammatical rules of index notation. In Minkowski coordinates (t, x, y, z) , this becomes $T^\mu_\mu \geq 0$, with the implied sum over μ expanding to give

$$T^t_t + T^x_x + T^y_y + T^z_z \geq 0.$$

The left-hand side of this relation, the sum of the main-diagonal elements of a matrix, is called the trace of the matrix, hence the name of this energy condition. Since this book uses the signature $+ - - -$ for the metric, raising the second index changes this to

$$T^{tt} - T^{xx} - T^{yy} - T^{zz} \geq 0.$$

In example 5 on p. 181, we computed the stress-energy tensor of a cloud of dust, in a frame moving at velocity v relative to the cloud's rest frame. The result was

$$T^{\mu'\nu'} = \begin{pmatrix} \gamma^2 \rho & \gamma^2 v \rho & 0 & 0 \\ \gamma^2 v \rho & \gamma^2 v^2 \rho & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

In this example, the trace energy condition is satisfied precisely under the condition $|v| \leq 1$, which can be interpreted as a statement that according the TEC, the mass-energy of the cloud can never be transported at a speed greater than c in any frame.

9.3 Gauss's theorem

9.3.1 Integral conservation laws

We've expressed conservation of charge and energy-momentum in terms of zero divergences,

$$\frac{\partial J^a}{\partial x^a} = 0$$

$$\frac{\partial T^{ab}}{\partial x^a} = 0.$$

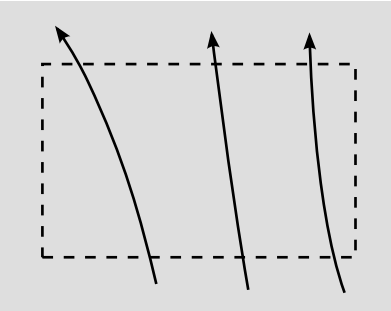
These are expressed in terms of derivatives. The derivative of a function at a certain point only depends on the behavior of the function near that point, so these are local statements of conservation. Conservation laws can also be stated globally: the total amount of something remains constant. Taking charge as an example, observer **o** defines Minkowski coordinates (t, x, y, z) , and at a time t_1 says that the total amount of charge in some region is

$$q(t_1) = \int_{t_1} J^a dS_a,$$

where the subscript t_1 means that the integrand is to be evaluated over the surface of simultaneity $t = t_1$, and $dS_a = (dx dy dz, 0, 0, 0)$ is an element of 3-volume expressed as a covector (p. 156). The charge at some later time t_2 would be given by a similar integral. If charge is conserved, and if our region is surrounded by an empty region through which no charge is coming in or out, then we should have $q(t_2) = q(t_1)$.

9.3.2 A simple form of Gauss's theorem

The connection between the local and global conservation laws is provided by a theorem called Gauss's theorem. In your course on electromagnetism, you learned Gauss's law, which relates the electric flux through a closed surface to the charge contained inside the surface. In the case where no charges are present, it says that the flux through such a surface cancels out. The interpretation is that since field lines only begin or end on charges, the absence of any charges means that the lines can't begin or end, and therefore, as in figure d, any field line that enters the surface (contributing some negative flux) must eventually come back out (creating some positive flux that cancels out the negative). But there is nothing about figure d that requires it to be interpreted as a drawing of electric field lines. It could just as easily be a drawing of the world-lines of some charged particles in $1 + 1$ dimensions. The bottom of the rectangle would then be the surface at t_1 and the top t_2 . We have $q(t_1) = 3$ and $q(t_2) = 3$ as well.



d / Three lines go in, and three come out. These could be field lines or world lines.

For simplicity, let's start with a very restricted version of Gauss's theorem. Let a vector field J^a be defined in two dimensions. (We

don't care whether the two dimensions are both spacelike or one spacelike and one timelike; that is, Gauss's theorem doesn't depend on the signature of the metric.) Let R be a rectangular area, and let S be its boundary. Define the flux of the field through S as

$$\Phi = \int_S J^a dS_a,$$

where the integral is to be taken over all four sides, and the covector dS_a points outward. If the field has zero divergence, $\partial J^a / \partial x^a = 0$, then the flux is zero.

Proof: Define coordinates x and y aligned with the rectangle. Along the top of the rectangle, the element of the surface, oriented outwards, is $d\mathbf{S} = (0, dx)$, so the contribution to the flux from the top is

$$\Phi_{\text{top}} = \int_{\text{top}} J^y(y_{\text{top}}) dx.$$

At the bottom, an outward orientation gives $d\mathbf{S} = (0, -dx)$, so

$$\Phi_{\text{bottom}} = - \int_{\text{bottom}} J^y(y_{\text{bottom}}) dx.$$

Using the fundamental theorem of calculus, the sum of these is

$$\Phi_{\text{top}} + \Phi_{\text{bottom}} = \int_R \frac{\partial J^y}{\partial y} dy dx.$$

Adding in the similar expressions for the left and right, we get

$$\Phi = \int_R \left(\frac{\partial J^x}{\partial x} + \frac{\partial J^y}{\partial y} \right) dx dy.$$

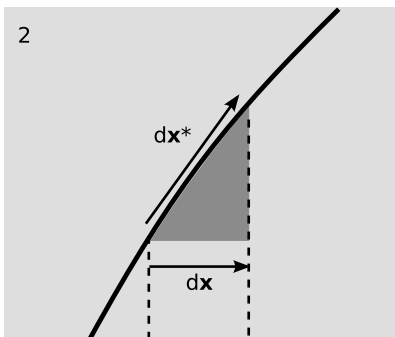
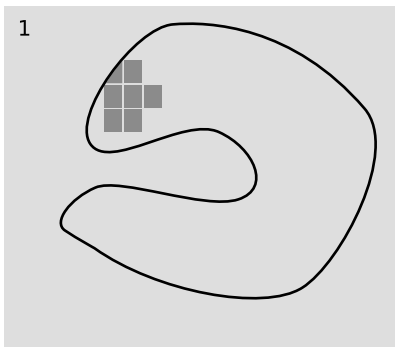
But the integrand is the divergence, which is zero by assumption, so $\Phi = 0$ as claimed.

9.3.3 The general form of Gauss's theorem

Although the coordinates were labeled x and y , the proof made no use of the metric, so the result is equally valid regardless of the signature. The rectangle could equally well have been a rectangle in $1 + 1$ -dimensional spacetime. The generalization to n dimensions is also automatic, and everything also carries through without modification if we replace the vector J^a with a tensor such as T^{ab} that has more indices — the extra index b just comes along for the ride. Sometimes, as with Gauss's law in electromagnetism, we are interested in fields whose divergences are not zero. Gauss's theorem then becomes

$$\int_S J^a dS_a = \int_R \frac{\partial J^a}{\partial x^a} dv,$$

where dv is the element of n -volume. In $3 + 1$ dimensions we could use Minkowski coordinates to write the element of 4-volume as $dv = dt dx dy dz$, and even though this expression is written in terms of



e / Proof of Gauss's theorem for a region with an arbitrary shape.

these specific coordinates, it is actually Lorentz invariant (section 2.5, p. 49).

The generalization to a region R with an arbitrary shape, figure e, is less trivial. The basic idea is to break up the region into rectangular boxes, $e/1$. Where the faces of two boxes coincide on the interior of R , their own outward directions are opposite. Therefore if we add up the fluxes through the surfaces of all the boxes, the contributions on the interior cancel, and we're left with only the exterior contributions. If R could be dissected exactly into boxes, then this would complete the proof, since the sum of exterior contributions would be the same as the flux through S , and the left-hand side of Gauss's theorem would be additive over the boxes, as is the right-hand side.

The difficulty arises because a smooth shape typically cannot be built out of bricks, a fact that is well known to Lego enthusiasts who build elaborate models of the Death Star. We could argue on physical grounds that no real-world measurement of the flux can depend on the granular structure of S at arbitrarily small scales, but this feels a little unsatisfying. For comparison, it is *not* strictly true that surface areas can be treated in this way. For example, if we approximate a unit 3-sphere using smaller and smaller boxes, the limit of the surface area is 6π , which is quite a bit greater than the surface area $4\pi/3$ of the limiting surface.

Instead, we explicitly consider the nonrectangular pieces at the surface, such as the one in e/2. In this drawing in $n = 2$ dimensions, the top of this piece is approximately a line, and in the limit we'll be considering, where its width becomes an infinitesimally small dx , the error incurred by approximating it as a line will be negligible. We define vectors dx and dx^* as shown in the figure. In more than the two dimensions shown in the figure, we would approximate the top surface as an $(n - 1)$ -dimensional parallelepiped spanned by vectors dx^* , dy^* , \dots . This is the point at which the use of the covector S_a (p. 156) pays off by greatly simplifying the proof.³ Applying this to the top of the triangle, dS is defined as the linear function that takes a vector \mathbf{J} and gives the n -volume spanned by \mathbf{J} along with dx^* , \dots .

Call the vertical coordinate on the diagram t , and consider the contribution to the flux from \mathbf{J} 's time component, J^t . Because the

³Here is an example of the ugly complications that occur if one doesn't have access to this piece of technology. In the low-tech approach, in Euclidean space, one defines an element of surface area $d\mathbf{A} = \hat{\mathbf{n}} dA$, where the unit vector $\hat{\mathbf{n}}$ is outward-directed with $\hat{\mathbf{n}} \cdot \hat{\mathbf{n}} = 1$. But in a signature such as $+- --$, we could have a region R such that over some large area of the bounding surface S , the normal direction was lightlike. It would therefore be impossible to scale $\hat{\mathbf{n}}$ so that $\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}$ was anything but zero. As an example of how much work it is to resolve such issues using stone-age tools, see Synge, *Relativity: The Special Theory*, VIII, §6-7, where the complete argument takes up 22 pages.

triangle's size is an infinitesimal of order dx , we can approximate J^t as being a constant throughout the triangle, while incurring only an error of order dx . (By stating Gauss's theorem in terms of derivatives of \mathbf{J} , we implicitly assumed it to be differentiable, so it is not possible for it to jump discontinuously.) Since $d\mathbf{S}$ depends linearly not just on \mathbf{J} but on all the vectors, the difference between the flux at the top and bottom of the triangle equals is proportional to the area spanned by \mathbf{J} and $d\mathbf{x}^* - d\mathbf{x}$. But the latter vector is in the t direction, and therefore the area it spans when taken with J^t is approximately zero. Therefore the contribution of J^t to the flux through the triangle is zero. To estimate the possible error due to the approximations, we have to count powers of dx . The possible variation of J^t over the triangle is of order $(dx)^1$. The covector $d\mathbf{S}$ is of order $(dx)^{n-1}$, so the possible error in the flux is of order $(dx)^n$.

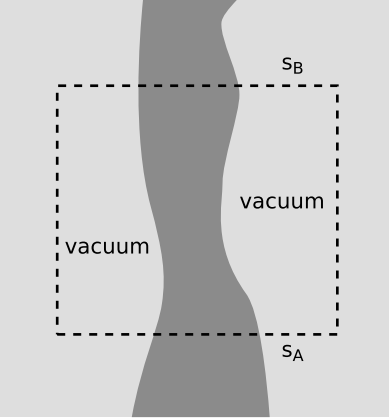
This was only an estimate of one part of the flux, the part contributed by the component J^t . However, we get the same estimate for the other parts. For example, if we refer to the two dimensions in figure e/2 as t and x , then interchanging the roles of t and x in the above argument produces the same error estimate for the contribution from J^x .

This is good. When we began this argument, we were motivated to be cautious by our observation that a quantity such as the surface area of R can't be calculated as the limit of the surface area as approximated using boxes. The reason we have that problem for surface area is that the error in the approximation on a small patch is of order $(dx)^{n-1}$, which is an infinitesimal of the same order as the surface area of the patch itself. Therefore when we scale down the boxes, the error doesn't get small compared to the total area. But when we consider flux, the error contributed by each of the irregularly shaped pieces near the surface goes like $(dx)^n$, which is of the order of the n -volume of the piece. This volume goes to zero in the limit where the boxes get small, and therefore the error goes to zero as well. This establishes the generalization of Gauss's theorem to a region R of arbitrary shape.

9.3.4 The energy-momentum vector

Einstein's celebrated $E = mc^2$ is a special case of the statement that energy-momentum is conserved, transforms like a four-vector, and has a norm m equal to the rest mass. Section 4.4 on p. 98 explored some of the problems with Einstein's original attempt at a proof of this statement, but only now are we prepared to completely resolve them. One of the problems was the definitional one of what we mean by the energy-momentum of a system that is not composed of pointlike particles. The answer is that for any phenomenon that carries energy-momentum, we must decide how it contributes to the stress-energy tensor. For example, the stress-energy tensor of the electric and magnetic fields is described in section 10.6 on p. 226.

For the reasons discussed in section 4.4 on p. 98, it is necessary to assume that energy-momentum is locally conserved, and also that the system being described is isolated. Local conservation is described by the zero-divergence property of the stress-energy tensor, $\partial T^{ab}/\partial x^a = 0$. Once we assume local conservation, figure f shows how to prove conservation of the integrated energy-momentum vector using Gauss's theorem. Fix a frame of reference \mathbf{o} . Surrounding the system, shown as a dark stream flowing through spacetime, we draw a box. The box is bounded on its past side by a surface that \mathbf{o} considers to be a surface of simultaneity s_A , and likewise on the future side s_B . It doesn't actually matter if the sides of the box are straight or curved according to \mathbf{o} . What does matter is that because the system is isolated, we have enough room so that between the system and the sides of the box there can be a region of vacuum, in which the stress-energy tensor vanishes.



f / Conservation of the integrated energy-momentum vector.

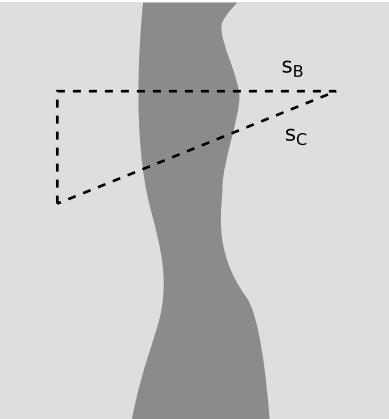
Observer \mathbf{o} says that at the initial time corresponding to s_A , the total amount of energy-momentum in the system was

$$p_A^\mu = - \int_{s_A} T^{\mu\nu} dS_\nu,$$

where the minus sign occurs because we take dS_ν to point outward, for compatibility with Gauss's theorem, and this makes it antiparallel to the velocity vector \mathbf{o} , which is the opposite of the orientation defined in equations (2) on p. 179. At the final time we have

$$p_B^\mu = \int_{s_B} T^{\mu\nu} dS_\nu,$$

with a plus sign because the outward direction is now the same as the direction of \mathbf{o} . Because of the vacuum region, there is no flux through the sides of the box, and therefore by Gauss's theorem $p_B^\mu - p_A^\mu = 0$. The energy-momentum vector has been globally conserved according to \mathbf{o} .



g / Lorentz transformation of the integrated energy-momentum vector.

We also need to show that the integrated energy-momentum transforms properly as a four-vector. To prove this, we apply Gauss's theorem to the region shown in figure g, where s_C is a surface of simultaneity according to some other observer \mathbf{o}' . Gauss's theorem tells us that $\mathbf{p}_B = \mathbf{p}_C$, which means that the energy-momentum on the two surfaces is the *same* vector in the absolute sense — but this doesn't mean that the two vectors have the same *components* as measured by different observers. Observer \mathbf{o} says that s_B is a surface of simultaneity, and therefore considers \mathbf{p}_B to be the total energy-momentum at a certain time. She says the total mass-energy is $p_B^\mu o_\mu$ (eq. (2a), p. 179), and similarly for the total momentum in the three spatial directions \mathbf{s}_1 , \mathbf{s}_2 , and \mathbf{s}_3 (eq. (2b)). Observer \mathbf{o}' , meanwhile, considers s_C to be a surface of simultaneity, and has the same interpretations for quantities such as $p_C^\mu o'_\mu$. But this is just a way of saying that \mathbf{p}_B^μ and \mathbf{p}_C^μ are related to each other by

a change of basis from $(\mathbf{o}, \mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3)$ to $(\mathbf{o}', \mathbf{s}'_1, \mathbf{s}'_2, \mathbf{s}'_3)$. A change of basis like this is just what we mean by a Lorentz transformation, so the integrated energy-momentum \mathbf{p} transforms as a four-vector.

9.3.5 Angular momentum

In sec. 8.2.2, p. 167, we gave physical and mathematical plausibility arguments for defining relativistic angular momentum as $L^{ab} = r^a p^b - r^b p^a$. We can now show that this quantity is actually conserved. Just as the flux of energy-momentum p^a is the stress-energy tensor T^{ab} , we can take the angular momentum L^{ab} and define its flux $\lambda^{abc} = r^a T^{bc} - r^b T^{ac}$. An observer with velocity vector o^c says that the density of energy-momentum is $T^{ac} o_c$ and the density of angular momentum is $\lambda^{abc} o_c$. If we can show that the divergence of λ with respect to its third index is zero, then it follows that angular momentum is conserved. The divergence is

$$\frac{\partial \lambda^{abc}}{\partial x^c} = \frac{\partial}{\partial x^c} (r^a T^{bc} - r^b T^{ac}).$$

The product rule gives

$$\frac{\partial \lambda^{abc}}{\partial x^c} = \delta_c^a T^{bc} + r^a \frac{\partial}{\partial x^c} T^{bc} - \delta_c^b T^{ac} - r^b \frac{\partial}{\partial x^c} T^{ac},$$

where δ_j^i , called the Kronecker delta, is defined as 1 if $i = j$ and 0 if $i \neq j$. The divergence of the stress-energy tensor is zero, so the second and fourth terms vanish, and

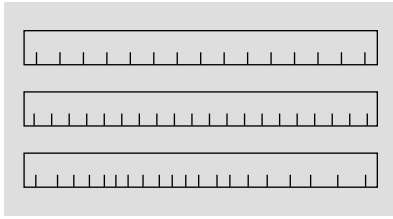
$$\begin{aligned} \frac{\partial \lambda^{abc}}{\partial x^c} &= \delta_c^a T^{bc} - \delta_c^b T^{ac} \\ &= T^{ba} - T^{ab}, \end{aligned}$$

but this is zero because the stress-energy tensor is symmetric.

9.4 ★ The covariant derivative

In this optional section we deal with the issues raised in section 7.5 on p. 150. We noted there that in non-Minkowski coordinates, one cannot naively use changes in the components of a vector as a measure of a change in the vector itself. A constant *scalar* function remains constant when expressed in a new coordinate system, but the same is not true for a constant vector function, or for any tensor of higher rank. This is because the change of coordinates changes the units in which the vector is measured, and if the change of coordinates is nonlinear, the units vary from point to point. This topic doesn't logically belong in this chapter, but I've placed it here because it can't be discussed clearly without already having covered tensors of rank higher than one.

Consider the one-dimensional case, in which a vector v^a has only one component, and the metric is also a single number, so that we



h / These three rulers represent three choices of coordinates.

can omit the indices and simply write v and g . (We just have to remember that v is really a vector, even though we're leaving out the upper index.) If v is constant, its derivative dv/dx , computed in the ordinary way without any correction term, is zero. If we further assume that the metric is simply the constant $g = 1$, then zero is not just the answer but the right answer.

Now suppose we transform into a new coordinate system X , and the metric G , expressed in this coordinate system, is not constant. Applying the tensor transformation law, we have $V = v dX/dx$, and differentiation with respect to X will not give zero, because the factor dX/dx isn't constant. This is the wrong answer: V isn't really varying, it just appears to vary because G does.

We want to add a correction term onto the derivative operator d/dX , forming a new derivative operator ∇_X that gives the right answer. ∇_X is called the covariant derivative. This correction term is easy to find if we consider what the result ought to be when differentiating the metric itself. In general, if a tensor appears to vary, it could vary either because it really does vary or because the metric varies. If the metric *itself* varies, it could be either because the metric really does vary or ... because the metric varies. In other words, there is no sensible way to assign a nonzero covariant derivative to the metric itself, so we must have $\nabla_X G = 0$. The required correction therefore consists of replacing d/dX with

$$\nabla_X = \frac{d}{dX} - G^{-1} \frac{dG}{dX}.$$

Applying this to G gives zero. G is a second-rank tensor with two lower indices. If we apply the same correction to the derivatives of other tensors of this type, we will get nonzero results, and they will be the right nonzero results.

Mathematically, the form of the derivative is $(1/y) dy/dx$, which is known as a logarithmic derivative, since it equals $d(\ln y)/dx$. It measures the *multiplicative* rate of change of y . For example, if y scales up by a factor of k when x increases by 1 unit, then the logarithmic derivative of y is $\ln k$. The logarithmic derivative of e^{cx} is c . The logarithmic nature of the correction term to ∇_X is a good thing, because it lets us take changes of scale, which are multiplicative changes, and convert them to additive corrections to the derivative operator. The additivity of the corrections is necessary if the result of a covariant derivative is to be a tensor, since tensors are additive creatures.

What about quantities that are not second-rank covariant tensors? Under a rescaling of coordinates by a factor of k , covectors scale by k^{-1} , and second-rank tensors with two lower indices scale by k^{-2} . The correction term should therefore be half as much for covectors,

$$\nabla_X = \frac{d}{dX} - \frac{1}{2} G^{-1} \frac{dG}{dX}.$$

and should have an opposite sign for vectors.

Generalizing the correction term to derivatives of vectors in more than one dimension, we should have something of this form:

$$\begin{aligned}\nabla_a v^b &= \partial_a v^b + \Gamma_{ac}^b v^c \\ \nabla_a v_b &= \partial_a v_b - \Gamma_{ba}^c v_c,\end{aligned}$$

where Γ_{ac}^b , called the Christoffel symbol, does not transform like a tensor, and involves derivatives of the metric. (“Christoffel” is pronounced “Krist-AWful,” with the accent on the middle syllable.)

An important gotcha is that when we evaluate a particular component of a covariant derivative such as $\nabla_2 v^3$, it is possible for the result to be nonzero even if the component v^3 vanishes identically.

Christoffel symbols on the globe

Example 10

As a qualitative example, consider the airplane trajectory shown in figure i, from London to Mexico City. This trajectory is the shortest one between these two points; such a minimum-length trajectory is called a geodesic. In physics it is customary to work with the colatitude, θ , measured down from the north pole, rather than the latitude, measured from the equator. At P, over the North Atlantic, the plane’s colatitude has a minimum. (We can see, without having to take it on faith from the figure, that such a minimum must occur. The easiest way to convince oneself of this is to consider a path that goes directly over the pole, at $\theta = 0$.)

At P, the plane’s velocity vector points directly west. At Q, over New England, its velocity has a large component to the south. Since the path is a geodesic and the plane has constant speed, the velocity vector is simply being parallel-transported; the vector’s covariant derivative is zero. Since we have $v_\theta = 0$ at P, the only way to explain the nonzero and positive value of $\partial_\phi v^\theta$ is that we have a nonzero and negative value of $\Gamma_{\phi\phi}^\theta$.

By symmetry, we can infer that $\Gamma_{\phi\phi}^\theta$ must have a positive value in the southern hemisphere, and must vanish at the equator.

$\Gamma_{\phi\phi}^\theta$ is computed in example 11 on page 197.

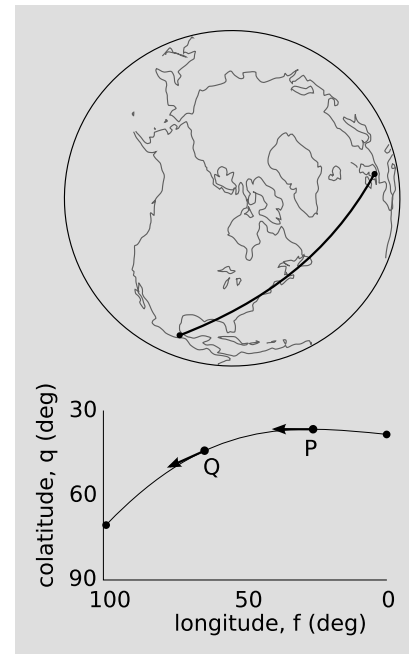
Symmetry also requires that this Christoffel symbol be independent of ϕ , and it must also be independent of the radius of the sphere.

To compute the covariant derivative of a higher-rank tensor, we just add more correction terms, e.g.,

$$\nabla_a U_{bc} = \partial_a U_{bc} - \Gamma_{ba}^d U_{dc} - \Gamma_{ca}^d U_{bd}$$

or

$$\nabla_a U_b^c = \partial_a U_b^c - \Gamma_{ba}^d U_d^c + \Gamma_{ad}^c U_b^d.$$



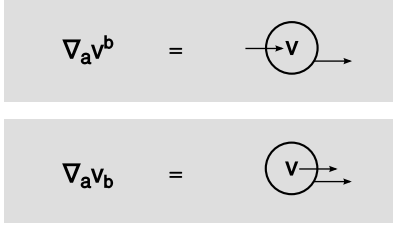
i / Example 10.

With the partial derivative ∂_μ , it does not make sense to use the metric to raise the index and form ∂^μ . It *does* make sense to do so with covariant derivatives, so $\nabla^a = g^{ab}\nabla_b$ is a correct identity.

9.4.1 Comma, semicolon, and birdtracks notation

Some authors use superscripts with commas and semicolons to indicate partial and covariant derivatives. The following equations give equivalent notations for the same derivatives:

$$\begin{aligned}\partial_\mu &= \frac{\partial}{\partial x^\mu} \\ \partial_\mu X_\nu &= X_{\nu,\mu} \\ \nabla_a X_b &= X_{b;a} \\ \nabla^a X_b &= X_b{}^{:a}\end{aligned}$$



j / Birdtracks notation for the covariant derivative.

Figure j shows two examples of the corresponding birdtracks notation. Because birdtracks are meant to be manifestly coordinate-independent, they do not have a way of expressing non-covariant derivatives.

9.4.2 Finding the Christoffel symbol from the metric

We've already found the Christoffel symbol in terms of the metric in one dimension. Expressing it in tensor notation, we have

$$\Gamma^d_{ba} = \frac{1}{2} g^{cd} (\partial_b g_{ca} + \partial_a g_{cb} - \partial_c g_{ab}),$$

where inversion of the one-component matrix G has been replaced by matrix inversion, and, more importantly, the question marks indicate that there would be more than one way to place the subscripts so that the result would be a grammatical tensor equation. The most general form for the Christoffel symbol would be

$$\Gamma^b_{ac} = \frac{1}{2} g^{db} (L \partial_c g_{ab} + M \partial_a g_{cb} + N \partial_b g_{ca}),$$

where L , M , and N are constants. Consistency with the one-dimensional expression requires $L + M + N = 1$. The condition $L = M$ arises on physical, not mathematical grounds; it reflects the fact that experiments have not shown evidence for an effect called torsion, in which vectors would rotate in a certain way when transported. The L and M terms have a different physical significance than the N term.

Suppose an observer uses coordinates such that all objects are described as lengthening over time, and the change of scale accumulated over one day is a factor of $k > 1$. This is described by the derivative $\partial_t g_{xx} < 1$, which affects the M term. Since the metric is used to calculate squared distances, the g_{xx} matrix element scales down by $1/\sqrt{k}$. To compensate for $\partial_t v^x < 0$, so we need to add a positive correction term, $M > 0$, to the covariant derivative. When

the same observer measures the rate of change of a vector v^t with *respect* to space, the rate of change comes out to be too *small*, because the variable she differentiates with respect to is too big. This requires $N < 0$, and the correction is of the same size as the M correction, so $|M| = |N|$. We find $L = M = -N = 1$.

Self-check: Does the above argument depend on the use of space for one coordinate and time for the other?

The resulting general expression for the Christoffel symbol in terms of the metric is

$$\Gamma_{ab}^c = \frac{1}{2} g^{cd} (\partial_a g_{bd} + \partial_b g_{ad} - \partial_d g_{ab}).$$

One can go back and check that this gives $\nabla_c g_{ab} = 0$.

Self-check: In the case of 1 dimension, show that this reduces to the earlier result of $-(1/2) dG/dX$.

Γ is not a tensor, i.e., it doesn't transform according to the tensor transformation rules. Since Γ isn't a tensor, it isn't obvious that the covariant derivative, which is constructed from it, *is* tensorial. But if it isn't obvious, neither is it surprising – the goal of the above derivation was to get results that would be coordinate-independent.

Christoffel symbols on the globe, quantitatively *Example 11*

In example 10 on page 195, we inferred the following properties for the Christoffel symbol $\Gamma_{\phi\phi}^\theta$ on a sphere of radius R : $\Gamma_{\phi\phi}^\theta$ is independent of ϕ and R , $\Gamma_{\phi\phi}^\theta < 0$ in the northern hemisphere (colatitude θ less than $\pi/2$), $\Gamma_{\phi\phi}^\theta = 0$ on the equator, and $\Gamma_{\phi\phi}^\theta > 0$ in the southern hemisphere.

The metric on a sphere is $ds^2 = R^2 d\theta^2 + R^2 \sin^2 \theta d\phi^2$. The only nonvanishing term in the expression for $\Gamma_{\phi\phi}^\theta$ is the one involving $\partial_\theta g_{\phi\phi} = 2R^2 \sin \theta \cos \theta$. The result is $\Gamma_{\phi\phi}^\theta = -\sin \theta \cos \theta$, which can be verified to have the properties claimed above.

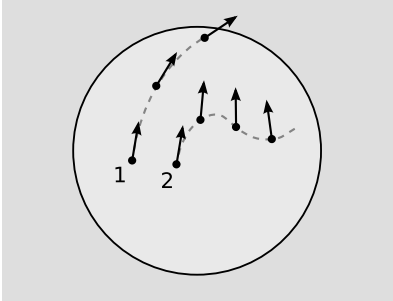
9.4.3 The geodesic equation

A world-line is a timelike curve in spacetime. As a special case, some such curves are actually not curved but straight. Physically, the ones we consider straight are those that could be the world-line of a test particle not acted on by any non-gravitational forces (sec. 5.1, p. 117). Mathematically, we will show in this section how the Christoffel symbols can be used to find differential equations that describe such motion. The world-line of a test particle is called a geodesic. The equations also have solutions that are spacelike or lightlike, and we consider these to be geodesics as well.

Geodesics play the same role in relativity that straight lines play in Euclidean geometry. In Euclidean geometry, we can specify two points and ask for the curve connecting them that has minimal length. The answer is a line. In special relativity, a timelike

geodesic *maximizes* the proper time (cf. section 2.4.2, p. 48) between two events.

In special relativity, geodesics are given by linear equations when expressed in Minkowski coordinates, and the velocity vector of a test particle has constant components when expressed in Minkowski coordinates. In general relativity, Minkowski coordinates don't exist, and geodesics don't have the properties we expect based on Euclidean intuition; for example, initially parallel geodesics may later converge or diverge.



k / The geodesic, 1, preserves tangency under parallel transport. The non-geodesic curve, 2, doesn't have this property; a vector initially tangent to the curve is no longer tangent to it when parallel-transported along it.

Characterization of the geodesic

A geodesic can be defined as a world-line that preserves tangency under parallel transport, k. This is essentially a mathematical way of expressing the notion that we have previously expressed more informally in terms of “staying on course” or moving “inertially.” (For reasons discussed in more detail on p. 200, this definition is preferable to defining a geodesic as a curve of extremal or stationary metric length.)

A curve can be specified by giving functions $x^i(\lambda)$ for its coordinates, where λ is a real parameter. A vector lying tangent to the curve can then be calculated using partial derivatives, $T^i = \partial x^i / \partial \lambda$. There are three ways in which a vector function of λ could change: (1) it could change for the trivial reason that the metric is changing, so that its components changed when expressed in the new metric; (2) it could change its components perpendicular to the curve; or (3) it could change its component parallel to the curve. Possibility 1 should not really be considered a change at all, and the definition of the covariant derivative is specifically designed to be insensitive to this kind of thing. 2 cannot apply to T^i , which is tangent by construction. It would therefore be convenient if T^i happened to be always the same length. If so, then 3 would not happen either, and we could reexpress the definition of a geodesic by saying that the covariant derivative of T^i was zero. For this reason, we will assume for the remainder of this section that the parametrization of the curve has this property. In a Newtonian context, we could imagine the x^i to be purely spatial coordinates, and λ to be a universal time coordinate. We would then interpret T^i as the velocity, and the restriction would be to a parametrization describing motion with constant speed. In relativity, the restriction is that λ must be an affine parameter. For example, it could be the proper time of a particle, if the curve in question is timelike.

Covariant derivative with respect to a parameter

The notation of section 9.4 is not quite adapted to our present purposes, since it allows us to express a covariant derivative with respect to one of the coordinates, but not with respect to a parameter such as λ . We would like to notate the covariant derivative of

T^i with respect to λ as $\nabla_\lambda T^i$, even though λ isn't a coordinate. To connect the two types of derivatives, we can use a total derivative. To make the idea clear, here is how we calculate a total derivative for a scalar function $f(x, y)$, without tensor notation:

$$\frac{df}{d\lambda} = \frac{\partial f}{\partial x} \frac{dx}{d\lambda} + \frac{\partial f}{\partial y} \frac{dy}{d\lambda}.$$

This is just the generalization of the chain rule to a function of two variables. For example, if λ represents time and f temperature, then this would tell us the rate of change of the temperature as a thermometer was carried through space. Applying this to the present problem, we express the total covariant derivative as

$$\begin{aligned}\nabla_\lambda T^i &= (\nabla_b T^i) \frac{dx^b}{d\lambda} \\ &= (\partial_b T^i + \Gamma^i_{bc} T^c) \frac{dx^b}{d\lambda}.\end{aligned}$$

The geodesic equation

Recognizing $\partial_b T^i dx^b/d\lambda$ as a total non-covariant derivative, we find

$$\nabla_\lambda T^i = \frac{dT^i}{d\lambda} + \Gamma^i_{bc} T^c \frac{dx^b}{d\lambda}.$$

Substituting $\partial x^i/\partial\lambda$ for T^i , and setting the covariant derivative equal to zero, we obtain

$$\frac{d^2 x^i}{d\lambda^2} + \Gamma^i_{bc} \frac{dx^c}{d\lambda} \frac{dx^b}{d\lambda} = 0.$$

This is known as the geodesic equation.

If this differential equation is satisfied for one affine parameter λ , then it is also satisfied for any other affine parameter $\lambda' = a\lambda + b$, where a and b are constants (problem 5, p. 214). Recall that affine parameters are only defined along geodesics, not along arbitrary curves. We can't start by defining an affine parameter and then use it to find geodesics using this equation, because we can't define an affine parameter without *first* specifying a geodesic. Likewise, we can't do the geodesic first and then the affine parameter, because if we already had a geodesic in hand, we wouldn't need the differential equation in order to find a geodesic. The solution to this chicken-and-egg conundrum is to write down the differential equations and try to find a solution, without trying to specify either the affine parameter or the geodesic in advance.

The geodesic equation is useful in establishing one of the necessary theoretical foundations of relativity, which is the uniqueness of geodesics for a given set of initial conditions. If the geodesic were not uniquely determined, then particles would have no way of deciding how to move. The form of the geodesic equation guarantees uniqueness, because one can use it to define an algorithm that constructs a geodesic for a given set of initial conditions.

Not characterizable as curves of stationary length

The geodesic equation may seem cumbersome. Why not just define a geodesic as a curve connecting two points that maximizes or minimizes its own metric length? The trouble is that this doesn't generalize nicely to curves that are not timelike. The casual reader may wish to skip the remainder of this subsection, which discusses this point.

For the spacelike case, we would want to define the proper metric length σ of a curve as $\sigma = \int \sqrt{-g_{ij}dx^i dx^j}$, the minus sign being necessary because we are using a metric with signature $+- - -$, and we want the result to be real. The quantity σ can be thought of as the result we would get by approximating the curve with a chain of short line segments, and adding their proper lengths. In the case where the whole curve lies within a plane of simultaneity for some observer, σ is the curve's Euclidean length as measured by that observer. Our σ is neither a maximum nor a minimum for a spacelike geodesic connecting two events. To see this, pick a frame in which the two events are simultaneous, and adopt Minkowski coordinates such that the points both lie on the x axis. Deforming the geodesic in the xy plane does what we expect according to Euclidean geometry: it increases the length. Deforming it in the xt plane, however, reduces the length (as becomes obvious when you consider the case of a large deformation that turns the geodesic into a curve of length zero, consisting of two lightlike line segments). The result is that the geodesic is neither a minimizer nor a maximizer of σ .

Maximizing or minimizing the proper length is a strong requirement. A related but more permissive criterion to apply to a curve connecting two fixed points is that if we vary the curve by some small amount, the variation in length should vanish to first order. For example, two points A and B on the surface of the earth determine a great circle, i.e., a circle whose circumference equals that of the earth. This great circle gives us two different paths by which we could travel from A to B. One of these will usually be longer than the other. Both of these are as straight as they can be while keeping to the surface of the earth, so in this context of spherical geometry they are both considered to be geodesics. One thing that the two paths have in common is that they are both *stationary*. Stationarity is defined as follows. Given a certain parametrized curve $\gamma(t)$, let us fix some vector $\mathbf{h}(t)$ at each point on the curve that is tangent to the earth's surface, and let \mathbf{h} be a continuous function of t that vanishes at the end-points. Then if ϵ is small compared to the radius of the earth, we can clearly define what it means to perturb γ by $\epsilon\mathbf{h}$, producing another curve γ^* similar to, but not the same as, γ . Stationarity means that the difference in length between γ and γ^* is of order ϵ^2 for small ϵ . This is a generalization of the elementary calculus notion that a function has a zero derivative near an extremum or point of inflection. In our example on the surface of the earth,

the two geodesics connecting A and B are both stationary.

Spacelike geodesics in special relativity are stationary by the above definition. However, this assertion may be misleading. Because we construct the displacement as the product $\epsilon \mathbf{h}$, its *derivative* is also guaranteed to shrink in proportion to ϵ for small ϵ . We could loosen this requirement a little bit, and only require that the magnitude of the displacement be of order ϵ . In this case, one can show that spacelike curves are not stationary. For example, any spacelike curve can be approximated to an arbitrary degree of precision by a chain of lightlike geodesic segments. Thus an arbitrarily small perturbation in the curve reduces its length to zero.

The situation becomes even worse for lightlike geodesics. Here we would have to define what “length” was. We could either take an absolute value, $L = \int \sqrt{|g_{ij}dx^i dx^j|}$, or not, $L = \int \sqrt{g_{ij}dx^i dx^j}$. If we don’t take the absolute value, L need not be real for small variations of the geodesic, and therefore we don’t have a well-defined ordering, and can’t say whether L is a maximum, a minimum, or neither. Regardless of whether we take the absolute value, we have $L = 0$ for a lightlike geodesic, but the square root function doesn’t have differentiable behavior when its argument is zero, so we don’t have stationarity. If we do take the absolute value, then for the geodesic curve, the length is zero, which is the shortest possible. However, one can have nongeodesic curves of zero length, such as a lightlike helical curve about the t axis.

9.5 ★ Congruences, expansion, and rigidity

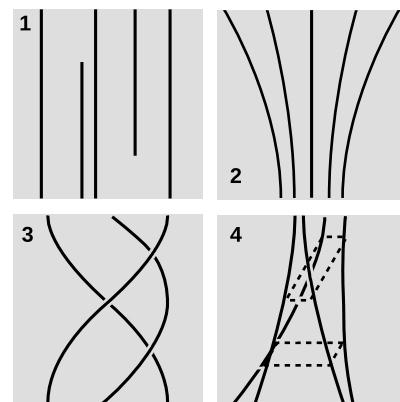
This chapter has focused on fluxes of conserved quantities; we wanted to rule out pictures like 1/1, in which the appearance and disappearance of world-lines would imply nonconservation of properties such as charge and mass-energy. But the mathematical techniques we’ve developed turn out to be an elegant way to approach the different issues described in the other parts of figure 1.

9.5.1 Congruences

In 1/2, we have *expansion*. For example, the world-lines could represent galaxies getting farther apart because of cosmological expansion resulting from the Big Bang. We do *not* expect rulers to expand or contract, in the sense that although a ruler may exhibit Lorentz contraction, it should always have the same length in its own rest frame unless it has been mechanically stressed or altered.

If there is more than one spatial dimension, then we can have *rotation*, as in 1/3. These world-lines could represent a constellation of orbiting satellites, or fixed points in a rotating laboratory.

The other interesting possibility, if there is more than one spatial dimension, is *shear*, figure 1/4. Here the rectangular group of four



1/1. Nonconservation. 2. Expansion. 3. Rotation. 4. Shear.

particles contracts in one direction while expanding in the other so as to keep the enclosed 2-volume constant.

In order to discuss these possibilities, it is convenient to define the notion of a timelike congruence, which is a set of nonintersecting, smooth, timelike world-lines whose union constitutes all the events in some region of spacetime. That is, we “fill in” spacetime with an infinite number of world-lines so that there is no space between them. This is something like the grain in a piece of wood, or Faraday’s conception of field lines filling space, except that one of our $n + 1$ dimensions is timelike, and the lines aren’t allowed to point in directions that lie outside the light cone. One way to specify a congruence is to give the normalized velocity vector that is tangent to the world-line passing through any given point.

An expanding congruence

Example 12

As an example of a congruence in $1 + 1$ dimensions, consider the set of all curves of the form $x = ae^{bt}$, where a and b are positive constants. It would look like figure l/2. Letting $u = dx/dt = abe^{bt}$, the velocity vector is $v^\lambda = \gamma^{-1}(1, u)$, where the factor of $\gamma^{-1} = \sqrt{1 - u^2}$ gives the proper normalization $v^\lambda v_\lambda = 1$.

A boring congruence

Example 13

Suppose we instead let the congruence consist of the set of all curves of the form $x = c + ut$, where c and u are constants and $|u| < 1$. Then as in example 12, $v^\lambda = \gamma^{-1}(1, u)$. The world-lines are inertial and parallel to one another.

9.5.2 Expansion and rigidity

For the remainder of this discussion, we restrict ourselves to the $1 + 1$ -dimensional case, so that rotation and shear are impossible, and the only interesting question is whether a given congruence has expansion. In $1 + 1$ dimensions, the congruence can be specified by giving the function $u(x, t)$, where as in examples 12 and 13, $u = dx/dt$. If u is constant, then we have example 13, and clearly there is no expansion. Thus expansion requires either $\partial u/\partial t$ or $\partial u/\partial x$, or both, to be nonzero.

m / 1. A congruence with $\partial u/\partial x$ equal to zero. 2. A congruence with $\partial u/\partial t = 0$. 3. A congruence without expansion.

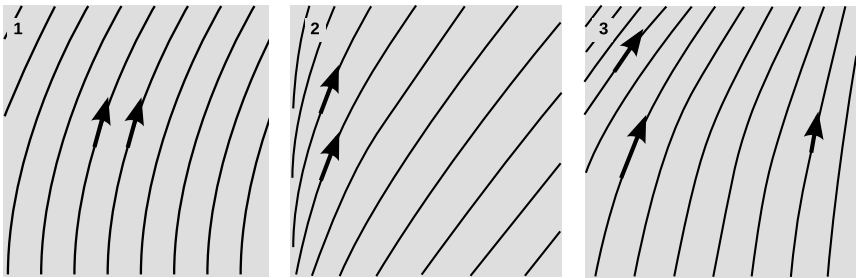


Figure m/1 shows the case where $\partial u/\partial x = 0$ and $\partial u/\partial t \neq 0$. Each world-line is a copy of the others that has been shifted spatially,

and the two velocity vectors shown as arrows are equal. This is precisely Bell's spaceship paradox (section 3.9.2, p. 71). Although the horizontal spacing between the world-lines remains constant as defined by the fixed frame of reference used for the diagram, an observer accelerating along with one of the particles would find that they had expanded away from one another, because the observer's meter-sticks have Lorentz-contracted. This is a real expansion in the sense that if the world-lines are particles in a solid object, the object comes under increasing tension.

In m/2 we have $\partial u/\partial t = 0$ and $\partial u/\partial x \neq 0$. The world-lines are copies of one another that have been shifted temporally. The two velocity vectors in the diagram are the same. All of the particles began accelerating from the same point in space, but at different times. Here there is clearly an expansion, because the world-lines are getting farther apart.

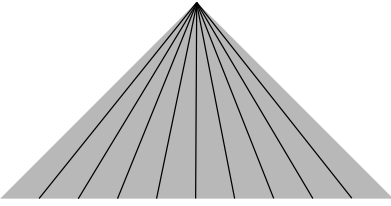
Suppose that we accelerate a rigid object such as a ruler. Then we must have something like m/3. To avoid the situations described in m/1 and m/2, the velocity vector must vary with both t and x ; the three velocity vectors in the figure are all different. As the particles accelerate, the spacing between them Lorentz-contracts, so that an observer accelerating along with them sees the spacing as remaining constant.

This notion of rigid motion in relativity is called Born rigidity. No physical substance can naturally be perfectly rigid (Born rigid), for if it were, then the speed at which sound waves traveled in it would be greater than c . Born rigidity can only be accomplished through a set of external forces applied at all points on the object according to a program that has been planned in advance. A real object such as a ruler does not maintain its own Born-rigidity, but it will eventually return to its original size and shape after having undergone relativistic acceleration, due to its own elastic properties, provided that the acceleration has been gentle enough to avoid permanently damaging it. In 1+1 dimensions, Born rigidity is equivalent to a lack of expansion. In 3 + 1 dimensions, we also require vanishing shear.

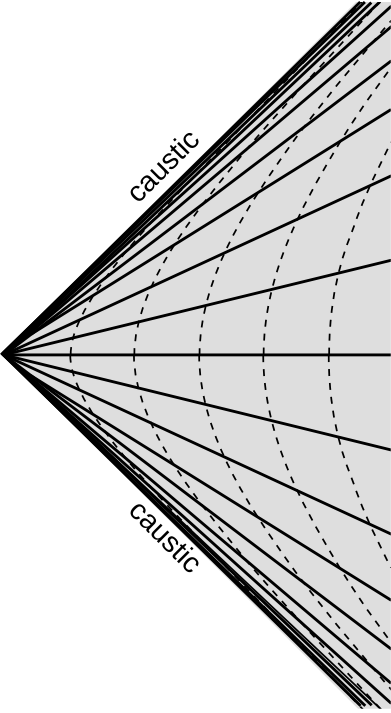
Mathematically, it is clear that the condition of vanishing expansion must be expressible in 1 + 1 dimensions in terms of the partial derivatives $\partial u/\partial t$ and $\partial u/\partial x$, and since we have been able to describe the condition in a frame-independent way (by referring it to observations made by the comoving observer), it should also be something we can express as a scalar within the grammar of index gymnastics. There is only one possible way to express such a condition, which is

$$\partial_a v^a = 0.$$

We can in fact define a scalar Θ , called the expansion scalar, ac-



n / Example 14.



o / A caustic in the lines of simultaneity of the family of accelerated world-lines.

cording to

$$\Theta = \partial_a v^a.$$

This definition is valid in $n + 1$ dimensions, but in $1 + 1$ dimensions it reduces to $\Theta = \partial\gamma/\partial t + \partial(u\gamma)/\partial x$.

The expansion scalar is interpreted as the fractional rate of change in the volume of a set of particles that move along the world-lines defined by the congruence, where the rate of change is defined with respect to the proper time τ of an observer moving along with the particles. For example, cosmological expansion leads to a fractional increase in the distances between galaxies $\Delta L/L$ which, for a small time interval $\Delta\tau$, is equal to $H_o\Delta\tau$, where H_o , called the Hubble constant, is about $2.3 \times 10^{-18} \text{ s}^{-1}$. That is, the fractional rate of change is $(1/L) dL/d\tau = H_o$. Because distances expand in all three spatial dimensions, the fractional rate of change of volume is $\Theta = (1/V) dV/d\tau = 3H_o$. (In this example, spacetime is not flat, so we would have to express Θ in terms of the covariant derivative ∇_a defined in section 9.4, not the partial derivative ∂_a .)

A catastrophe

Example 14

Consider the timelike congruence in $1 + 1$ dimensions defined by $u = x/t$. This consists of the set of all inertial world-lines passing through the origin. Since our definition of a congruence requires that the world-lines be non-intersecting, let's restrict this example to the interior of the past light cone of the origin, $|x| < -t$. We have a universe full of hapless particles, all heading like lemmings toward a catastrophic collision. The spacetime diagram looks like an optical ray diagram for the formation of a real image. A computation gives the unexpectedly simple result $\Theta = \gamma/t$. For $t < 0$, this is negative, indicating a contraction, and it blows up to minus infinity as t approaches 0.

9.5.3 Caustics

The apex of the cone in example 14 is a *caustic*. Given a space-filling set of straight lines, a caustic occurs where their intensity diverges to infinity. The word means “burning,” because in optics a caustic of light rays concentrates energy and can burn things. Example 14 involves a caustic of timelike world-lines, and “straight” is to be interpreted as meaning that the world-lines are inertial.

Figure o shows two caustics formed by spacelike lines for the accelerated coordinate system described in 7.1. Here, as is often the case, the caustics are not just points.

An example from general relativity is that when a black hole forms by gravitational collapse, a caustic is formed at a one point by the set of lightlike world-lines that enter the event horizon from the outside universe at the moment when the horizon is formed. If a ray of light is emitted from this caustic point, it remains on the event horizon forever, as do all rays emitted at the horizon in the

outward direction at later times. The event horizon is the same set of events as the union of all the lightlike world-lines that enter the horizon at the caustic.⁴

9.5.4 The Herglotz-Noether theorem in 1+1 dimensions

Certain Born-rigid types of motion are possible, and others are not, purely as a matter of kinematics. It turns out to be possible to accelerate a rod in a Born-rigid way along its own length (problem 7, p. 214), but surprisingly, it is not possible, for example, for a sphere to remain Born-rigid while simultaneously rotating and having its center of mass accelerated. The possible types of motion are delineated by a theorem called the Herglotz-Noether theorem. Unlike the 3 + 1-dimensional version of the theorem, the 1 + 1-dimensional version is neither surprising nor difficult to state or prove.

Herglotz-Noether theorem in 1+1 dimensions: Any rigid motion in 1 + 1 dimensions is uniquely determined by the world-line W of one point, provided that the world-line of that point is smooth and timelike. It is in general only possible to extend the congruence describing the motion to some neighborhood of W .

Proof: To avoid technical issues, we assume that “smooth” means analytic, which slightly weakens the result. As discussed above, zero expansion is equivalent to $0 = \partial\gamma/\partial t + \partial(u\gamma)/\partial x$, where (t, x) are any set of Minkowski coordinates. This can be put in the form $\partial u/\partial x = f(u)\partial u/\partial t$, where f is smooth for $-1 < u < 1$ and $f(0) = 0$. We need to prove that the solution of this partial differential equation, if it exists, is unique given W . We arbitrarily choose one event on W . By assumption, W is timelike at this point, so we are free to choose our Minkowski coordinates such that our point is at rest at this event at the origin. Since $f(0) = 0$, it follows that at the origin $\partial u/\partial x = 0$. We can similarly evaluate the higher derivatives $\partial^n u/\partial x^n$, and because u is smooth we can in this calculation freely interchange the order of the partial derivatives ∂_x and ∂_t . It is straightforward to show that these higher derivatives $\partial^n u/\partial x^n$ are also zero. Since $u(x)$ is assumed to be analytic, it follows that $u(0, x) = 0$ for all x , i.e., an observer instantaneously moving along W at $t = 0$ says that all other points are at rest as well at that time. But because W is timelike, we can always find some neighborhood A of W such that every point in A is simultaneous with a unique event on W according to an observer at that event moving along with W . (Cf. p. 73.) Therefore the value of u is determined everywhere in A , and this completes the proof that the congruence exists and is unique in A .

Remarks: (1) The 1 + 1-dimensional version of the Herglotz-Noether theorem is not a special case of the 3 + 1-dimensional ver-

⁴Penrose, 1968. The proof is presented in Misner, Thorne, and Wheeler, p. 924

sion. The latter is usually proved for a space-filling congruence, and it fails when the body in question does not enclose a volume, e.g., in the case of a thin rod or a letter “C.”

(2) The theorem can be strengthened by relaxing the requirement of smoothness so that only the existence of a second derivative of the position with respect to proper time is required.⁵

(3) If the motion is accelerated, then the rigid motion cannot be extended to an arbitrary distance from W . If the proper acceleration of W can be as great as a , then as in example ??, p. ??, we expect to be able to extend the rigid motion to a proper distance only as big as c^2/a , where there will be a caustic similar to the one in figure o.

9.5.5 Bell’s spaceship paradox revisited

Bell’s spaceship paradox was discussed in section 3.9.2 on p. 71. In the paradox, two spaceships begin accelerating simultaneously and have equal accelerations in the frame of an external, inertial observer, causing a thread stretched between them breaks. We now give a more rigorous and mathematically elegant demonstration of the same result, suggested by P. Allen.

The motion of the thread throughout its length can be described by a timelike congruence. If the thread is not to come under any strain, then this must be a Born-rigid congruence. By the 1 + 1-dimensional Herglotz-Noether theorem, the congruence is uniquely determined by the motion of one of its points, which we take to be the trailing rocket. This congruence happens to be known. It is defined by the system of accelerated coordinates (Rindler coordinates) described in section 7.1, p. 143. The vanishing of the expansion scalar for this congruence is left for the reader to verify (problem 7, p. 214). But this congruence consists of world-lines whose proper accelerations are each constant and all different from one another, and this is inconsistent with the description given in the Bell paradox, where it is stated that a frame exists in which the motions of the two ships are identical except for a translation. Therefore the thread cannot move rigidly.

This completes the resolution fo the paradox, but as an illustrative example, we present an explicit calculation of the expansion scalar for the congruence that one would most naturally imagine to be implied by the description of the paradox. This is given by $(x + c)^2 = 1 + t^2$. For a given value of the parameter c , we get an accelerating world-line. (Its proper acceleration $a = 1$ happens to be constant, example 4, p. 61, although this is not necessary for the purposes of discussing the paradox.) Each world-line starts at rest at $t = 0$, and each one has the same acceleration at any given t . By

⁵Giulini, “The Rich Structure of Minkowski Space,” arxiv.org/abs/0802.4345, theorems 18 and 22

picking any two distinct values of c as the endpoints of the thread, we obtain the literal situation described in the paradox.

Implicit differentiation gives $u = t/\sqrt{1+t^2}$. The algebra gets a little messy now, so I used the open-source computer algebra system Maxima. The following program, which should be fairly readable without previous knowledge of Maxima's syntax, calculates the expansion tensor:

```
1  u:t/sqrt(1+t^2);
2  gamma:1/sqrt(1-u^2);
3  theta:diff(gamma,t)+diff(u*gamma,x);
4  is(equal(theta,gamma*u^2/t));
```

The third line prints out a complicated expression for Θ , which the fourth line shows can be simplified to $\gamma u^2/t$. This is positive for $t > 0$, which shows that the thread is forced to expand. Note that although the calculation was carried out in a particular set of coordinates, a relativistic scalar such as Θ has a coordinate-independent value. Reference to a particular coordinate system or frame of reference occurs only in the initial definition of the congruence, which is defined in order to model the situation described in the paradox, which is stated in terms of a particular external observer.

9.6 Units of measurement for tensors

Analyzing units, also known as dimensional analysis, is one of the first things we learn in freshman physics. It's a useful way of checking our math, and it seems as though it ought to be straightforward to extend the technique to relativity. It certainly can be done, but it isn't quite as trivial as might be imagined. We'll see below that different authors prefer differing systems, and clashes occur between some of the notational systems in use.

One of our most common jobs is to change from one set of units to another, but in relativity it becomes nontrivial to define what we mean by the notion that our units of measurement change or don't change. We could, e.g., appeal to an atomic standard, but Dicke⁶ points out that this could be problematic. Imagine, he says, that

you are told by a space traveller that a hydrogen atom on Sirius has the same diameter as one on the earth. A few moments' thought will convince you that the statement is either a definition or else meaningless.

(Some related ideas about the numerical value of c were discussed on p. 19.)

⁶“Mach's principle and invariance under transformation of units,” *Phys Rev* 125 (1962) 2163

To start with, we note that abstract index notation is more convenient than concrete index notation for these purposes. As noted in section 7.5, p. 150, concrete index notation assigns different units to different components of a tensor if we use coordinates, such as spherical coordinates (t, r, θ, ϕ) , that don't all have units of length. In abstract index notation, a symbol like v^i stands for the whole vector, not for one of its components. Since abstract index notation does not even offer us a notation for components, if we want to apply dimensional analysis we must define a system in which units are attributed to a tensor as a whole. Suppose we write down the abstract-index form of the equation for proper time:

$$ds^2 = g_{ab} dx^a dx^a$$

In abstract index notation, dx^a doesn't mean an infinitesimal change in a particular coordinate, it means an infinitesimal displacement vector.⁷ This equation has one quantity on the left and three factors on the right. Suppose we assign these parts of the equation units $[ds] = L^\sigma$, $[g_{ab}] = L^{2\gamma}$, and $[dx^a] = [dx^b] = L^\xi$, where square brackets mean “the units of” and L stands for units of length. We then have $\sigma = \gamma + \xi$. Due to the ambiguities referred to above, we can pick any values we like for these three constants, as long as they obey this rule. I find $(\sigma, \gamma, \xi) = (1, 0, 1)$ to be natural and convenient, but Dicke, in the above-referenced paper, likes $(1, 1, 0)$, while the mathematician Terry Tao advocates $(0, \mp 1, \pm 1)$.

Suppose we raise and lower indices to form a tensor with r upper indices and s lower indices. We refer to this as a tensor of rank (r, s) . (We don't count contracted indices, e.g., $u^a v_a$ is a rank- $(0, 0)$ scalar.) Since the metric is the tool we use for raising and lowering indices, and the units of the lower-index form of the metric are $L^{2\gamma}$, it follows that the units vary in proportion to $L^{\gamma(s-r)}$. In general, you can assign a physical quantity units L^u that are a product of two factors, a “kinematical” or purely geometrical factor L^k , where $k = \gamma(s - r)$, and a dynamical factor $L^d \dots$, which can depend on what kind of quantity it is, and where the \dots indicates that if your system of units has more than just one base unit, those can be in there as well. Dicke uses units with $\hbar = c = 1$, for example, so there is only one base unit, and mass has units of inverse length and $d_{\text{mass}} = -1$. In general relativity it would be more common to use units in which $G = c = 1$, which instead give $d_{\text{mass}} = +1$.

The units of momentum

Example 15

Consider the equation

$$p^a = mv^a$$

for the momentum of a material particle. Suppose we use special-relativistic units in which $c = 1$, but because gravity isn't incorpo-

⁷For a modern and rigorous development of differential geometry along these lines, see Nowik and Katz, arxiv.org/abs/1405.0984.

rated into the theory, G plays no special role, and it is natural to use a system of units in which there is a base unit of mass M .

The kinematic units check out, because $k_p = k_m + k_v$:

$$\gamma(-1) = \gamma(0) + \gamma(-1)$$

This is merely a matter of counting indices, and was guaranteed to check out as long as the indices were written in a grammatical way on both sides of the equation. What this check is essentially telling us is that if we were to establish Minkowski coordinates in a neighborhood of some point, and do a change of coordinates $(t, x, y, z) \rightarrow (\alpha t, \alpha x, \alpha y, \alpha z)$, then the quantities on both sides of the equation would vary under the tensor transformation laws according to the same exponent of α . For example, if we changed from meters to centimeters, the equation would still remain valid.

For the dynamical units, suppose that we use $(\sigma, \gamma, \xi) = (1, 0, 1)$, so that an infinitesimal displacement dx^a has units of length L , as does proper time ds . These two quantities are purely kinematic, so we don't assign them any dynamical units, and therefore the velocity vector $v^a = dx^a/ds$ also has no dynamical units. Our choice of a system of units gives $[m] = M$. We require that the equation $p^a = mv^a$ have dynamical units that check out, so:

$$M = 1 \cdot M$$

We must also assign units of mass to the momentum.

A system almost identical to this one, but with different terminology, is given by Schouten.⁸

For practical purposes in checking the units of an equation, we can see from example 15 that worrying about the kinematic units is a waste of time as long as we have checked that the indices are grammatical. We can therefore give a simplified method that suffices for checking the units of any equation in abstract index notation.

1. We assign a tensor the same units that one of its concrete components *would* have if we were to adopt (local) Minkowski coordinates, in the system with $(\sigma, \gamma, \xi) = (1, 0, 1)$. These are the units we would automatically have imputed to it after learning special relativity but before learning about tensors or fancy coordinate transformations. Since $\gamma = 0$, the positions of the indices do not affect the result.
2. The units of a sum are the same as the units of the terms.
3. The units of a tensor product are the product of the units of the factors.

⁸Tensor Analysis for Physicists, ch. VI

9.7 ★ Notations for tensors

Johnny is an American grade-school kid who has had his tender mind protected from certain historical realities, such as the political status of slaves, women, and Native Americans in the early United States. If Johnny ever tries to read the U.S. Constitution, he will be very confused by certain passages, such as the infamous three-fifths clause referring opaquely to “all other persons.”

This optional section is meant to expose you to some similar historical ugliness involving tensor notation, knowledge of which may be helpful if you learn general relativity in the future. As in the evolution of the U.S. Constitution and its interpretation, we will find that not all the changes have been improvements. In sections 9.7.1-9.7.2 we briefly recapitulate some notations that have already been introduced, and then in sections 9.7.3-9.7.4 we introduce two new ones.

9.7.1 Concrete index notation

A displacement vector is our prototypical example of a tensor, and the original nineteenth-century approach was to associate this tensor with the changes in the coordinates. Tensors achieve their full importance in differential geometry, where space (or spacetime, in general relativity) may be curved, in the sense defined in section 2.2, p. 45. In this context, only infinitesimally small displacements qualify as vectors; to see this, imagine displacements on a sphere, which do not commute for the reasons described in section 8.3.1, p. 170. On small scales, the sphere’s curvature is not apparent, which is why we need to make our displacements infinitesimal. Thus in this approach, the simplest example of a relativistic tensor occurs if we pick Minkowski coordinates to describe a region of spacetime that is small enough for the curvature to be negligible, and we associate a displacement vector with a 4-tuple of infinitesimal changes in the coordinates:

$$(dt, dx, dy, dz)$$

Until about 1960, this carried the taint of the lack of rigor believed to be associated with Leibniz-style infinitesimal numbers, but this difficulty was resolved and is no longer an argument against the notation.⁹

9.7.2 Coordinate-independent notation

A more valid reason for disliking the old-school notation is that, as described in ch. 7, p. 143, it is desirable to avoid writing every line of mathematics in a notation that explicitly refers to a choice of coordinates. We might therefore prefer, as Penrose began advocating around 1970, to notate this vector in coordinate-independent

⁹For a thorough development of the “back-to-the-future” use of infinitesimals for this purpose, see Nowik and Katz, arxiv.org/abs/1405.0984.

notation such as “birdtracks” (section 6.1.3, p. 126),

$$\rightarrow d\mathbf{x},$$

or the synonymous abstract index notation (section ??, p. ??),

$$dx^a,$$

where the use of the Latin letter a means that we’re not referring to any coordinate system, a doesn’t take on values such as 1 or 2, and dx^a refers to the entire object $\rightarrow d\mathbf{x}$, not to some real number or set of real numbers.

Unfortunately for the struggling student of relativity, there are at least two more notations now in use, both of them incompatible in various ways with the ones we’ve encountered so far.

9.7.3 Cartan notation

Our notation involving upper and lower indices is descended from a similar-looking one invented in 1853 by Sylvester.¹⁰ In this system, vectors are thought of as invariant quantities. We write a vector in terms of a basis $\{\mathbf{e}_\mu\}$ as $\mathbf{x} = \sum x^\mu \mathbf{e}_\mu$. Since \mathbf{x} is considered invariant, it follows that the components x^μ and the basis vectors \mathbf{e}_μ must transform in opposite ways. For example, if we convert from meters to centimeters, the x^μ get a hundred times bigger, which is compensated for by a corresponding shrinking of the basis vectors by 1/100.

This notation clashes with normal index notation in certain ways. One gotcha is that we can’t infer the rank of an expression by counting indices. For example, $\mathbf{x} = \sum x^\mu \mathbf{e}_\mu$ is notated as if it were a scalar, but this is actually a notation for a vector.

Circa 1930, Élie Cartan augmented this notation with a trick that is perhaps a little too cute for its own good. He noted that the partial differentiation operators $\partial/\partial x^\mu$ could be used as a basis for a vector space whose structure is the same as the space of ordinary vectors. In the modern context we rewrite the operator $\partial/\partial x^\mu$ as ∂_μ and use the Einstein summation convention, so that in the Cartan notation we express a vector in terms of its components as

$$\mathbf{x} = x^\mu \partial_\mu.$$

In the Cartan notation, the symbol dx^μ is hijacked in order to represent something completely different than it normally does; it’s taken to mean the dual vector corresponding to ∂_μ . The set $\{dx^\mu\}$ is used as a basis for notating covectors.

A further problem with the Cartan notation arises when we try to use it for dimensional analysis (see section 9.7.5).

¹⁰An easily obtainable modern description is given in Coxeter, *Introduction to Geometry*.

9.7.4 Index-free notation

Independently of Penrose and the physics community, mathematicians invented a different coordinate-free notation, one without indices. In this notation, for example, we would notate the magnitude of a vector not as $v_a v^a$ or $g_{ab} v^a v^b$ but as

$$g(\mathbf{v}, \mathbf{v}).$$

This notation is too clumsy for use in complicated expressions involving tensors with many indices. As shown in section 9.7.5, it is also not compatible with the way physicists are accustomed to doing dimensional analysis.

9.7.5 Incompatibility of Cartan and index-free notation with dimensional analysis

In section 9.6 we developed a system of dimensional analysis for use with abstract index notation. Here we discuss the issues that arise when we attempt to mix in other notational systems.

One of the hallmarks of index-free notation is that it uses non-multiplicative notation for many tensor products that would have been written as multiplication in index notation, e.g., $g(\mathbf{v}, \mathbf{v})$ rather than $v^a v_a$. This makes the system clumsy to use for dimensional analysis, since we are accustomed to reasoning about units based on the assumption that the units of any term in an equation equal the product of the units of its factors.

In Cartan notation we have the problem that certain notations, such as dx^μ , are completely redefined. The remainder of this section is devoted to exploring what goes wrong when we attempt to extend the analysis of section 9.6 to include Cartan notation. Let vector \mathbf{r} and covector $\boldsymbol{\omega}$ be duals of each other, and let \mathbf{r} represent a displacement. In Cartan notation, we write these vectors in terms of their components, in some coordinate system, as follows:

$$\mathbf{r} = r^\mu \partial_\mu \quad (5)$$

$$\boldsymbol{\omega} = \omega_\mu dx^\mu \quad (6)$$

Suppose that the coordinates are Minkowski. Reading from left to right and from top to bottom, there are six quantities occurring in these equations. We attribute to them the units L^A, L^B, \dots, L^F . If we follow the rule that multiplicative notation is to imply multiplication of units, then

$$A = B + C \quad \text{and} \quad (7)$$

$$D = E + F. \quad (8)$$

For compatibility with the system in section 9.6, equations 5-6 require

$$A + D = 2\sigma \quad \text{and} \quad (9)$$

$$D = 2\gamma + B. \quad (10)$$

To avoid a clash between Cartan and concrete index notation in a Minkowski coordinate system, it would appear that we want the following three additional conditions.

$$F = \xi \quad \text{units of Cartan } dx^\mu \text{ not to clash with units of } dx^\mu \quad (11)$$

$$C = -\xi \quad \text{units of Cartan } \partial_\mu \text{ not to clash with units of the derivative} \quad (12)$$

$$B = \xi \quad \text{units of components in Cartan notation not to clash with units of } dx^\mu \quad (13)$$

We have 6 unknowns and 7 constraints, so in general Cartan notation cannot be incorporated into this system without some constraint on the exponents (σ, γ, ξ) . In particular, we require $\xi = 0$, which is not a choice that most physicists prefer.

Problems

1 Rewrite the stress-energy tensor of a perfect fluid in SI units. For air at sea level, compare the sizes of its components.

2 Prove by direct computation that if a rank-2 tensor is symmetric when expressed in one Minkowski frame, the symmetry is preserved under a boost.

3 Consider the following change of coordinates:

$$t' = -t$$

$$x' = x$$

$$y' = y$$

$$z' = z$$

This is called a time reversal. As in example 6 on p. 181, find the effect on the stress-energy tensor.

4 Show that in Minkowski coordinates in flat spacetime, all Christoffel symbols vanish.

5 Show that if the differential equation for geodesics on page 199 is satisfied for one affine parameter λ , then it is also satisfied for any other affine parameter $\lambda' = a\lambda + b$, where a and b are constants.

6 This problem investigates a notational conflict in the description of the metric tensor using index notation. Suppose that we have two different metrics, $g_{\mu\nu}$ and $g'_{\mu\nu}$. The difference of two rank-2 tensors is also a rank-2 tensor, so we would like the quantity $\delta g_{\mu\nu} = g'_{\mu\nu} - g_{\mu\nu}$ to be a well-behaved tensor both in its transformation properties and in its behavior when we manipulate its indices. Now we also have $g^{\mu\nu}$ and $g'^{\mu\nu}$, which are defined as the matrix inverses of their lower-index counterparts; this is a special property of the metric, not of rank-2 tensors in general. We can then define $\delta g^{\mu\nu} = g'^{\mu\nu} - g^{\mu\nu}$. (a) Use a simple example to show that $\delta g_{\mu\nu}$ and $\delta g^{\mu\nu}$ cannot be computed from one another in the usual way by raising and lowering indices. (b) Find the general relationship between $\delta g_{\mu\nu}$ and $\delta g^{\mu\nu}$.

7 In section 9.5.5 on p. 206, we analyzed the Bell spaceship paradox using the expansion scalar and the Herglotz-Noether theorem. Suppose that we carry out a similar analysis, but with the congruence defined by $x^2 - t^2 = a^{-2}$. The motivation for considering this congruence is that its world-lines have constant proper acceleration a , and each such world-line has a constant value of the coordinate X in the system of accelerated coordinates (Rindler coordinates) described in section 7.1, p. 143. Show that the expansion tensor vanishes. The interpretation is that it is possible to apply a carefully planned set of external forces to a straight rod so that it accelerates along its own length without any stress, i.e., while remaining Born-rigid.

Chapter 10

Electromagnetism

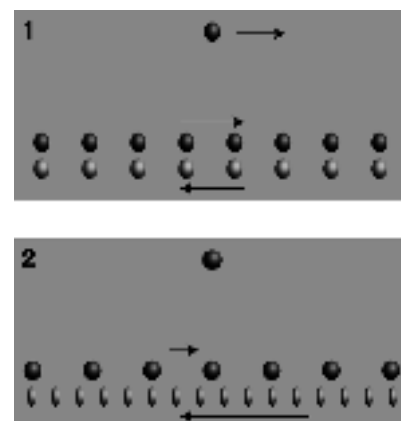
10.1 Relativity requires magnetism

Figure a/1 is an unrealistic model of charged particle moving parallel to a current-carrying wire. What electrical force does the lone particle in figure a/1 feel? Since the density of “traffic” on the two sides of the “road” is equal, there is zero overall electrical force on the lone particle. Each “car” that attracts the lone particle is paired with a partner on the other side of the road that repels it. If we didn’t know about magnetism, we’d think this was the whole story: the lone particle feels no force at all from the wire.

Figure a/2 shows what we’d see if we were observing all this from a frame of reference moving along with the lone charge. Relativity tells us that moving objects appear contracted to an observer who is not moving along with them. Both lines of charge are in motion in both frames of reference, but in frame 1 they were moving at equal speeds, so their contractions were equal. In frame 2, however, their speeds are unequal. The dark charges are moving more slowly than in frame 1, so in frame 2 they are less contracted. The light-colored charges are moving more quickly, so their contraction is greater now. The “cars” on the two sides of the “road” are no longer paired off, so the electrical forces on the lone particle no longer cancel out as they did in a/1. The lone particle is attracted to the wire, because the particles attracting it are more dense than the ones repelling it.

Now observers in frames 1 and 2 disagree about many things, but they do agree on concrete events. Observer 2 is going to see the lone particle drift toward the wire due to the wire’s electrical attraction, gradually speeding up, and eventually hit the wire. If 2 sees this collision, then 1 must as well. But 1 knows that the total electrical force on the lone particle is exactly zero. There must be some new type of force. She invents a name for it: magnetism.

Magnetism is a purely relativistic effect. Since relativistic effects are down by a factor of v^2 compared to Newtonian ones, it’s surprising that relativity can produce an effect as vigorous as the attraction between a magnet and your refrigerator. The explanation is that although matter is electrically neutral, the cancellation of electrical forces between macroscopic objects is extremely delicate, so anything that throws off the cancellation, even slightly, leads to a surprisingly large force.



a / A model of a charged particle and a current-carrying wire, seen in two different frames of reference. The relativistic length contraction is highly exaggerated. The force on the lone particle is purely magnetic in 1, and purely electric in 2.

10.2 Fields in relativity

Based on what we learned in section 10.1, the next natural step would seem to be to find some way of extending Coulomb's law to include magnetism. For example, we could try to find a formula for the magnetic force between charges q_1 and q_2 based on not just their relative positions but also on their velocities. The following considerations, however, tell us not to go down that path.

10.2.1 Time delays in forces exerted at a distance

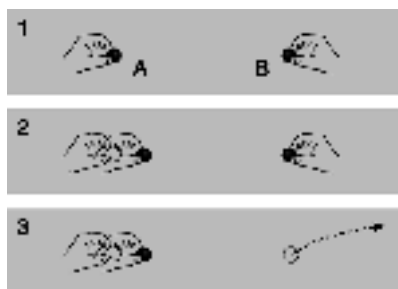
Relativity forbids Newton's instantaneous action at a distance (p. 17). Since forces can't be transmitted instantaneously, it becomes natural to imagine force-effects spreading outward from their source like ripples on a pond, and we then have no choice but to impute some physical reality to these ripples. We call them fields, and they have their own independent existence.

Even empty space, then, is not perfectly featureless. It has measurable properties. For example, we can drop a rock in order to measure the direction of the gravitational field, or use a magnetic compass to find the direction of the magnetic field. This concept made a deep impression on Einstein as a child. He recalled that as a five-year-old, the gift of a magnetic compass convinced him that there was "something behind things, something deeply hidden."

10.2.2 Fields carry energy.

The smoking-gun argument for this strange notion of traveling force ripples comes from the fact that they carry energy. In figure b/1, Alice and Betty hold positive charges A and B at some distance from one another. If Alice chooses to move her charge closer to Betty's, b/2, Alice will have to do some mechanical work against the electrical repulsion, burning off some of the calories from that chocolate cheesecake she had at lunch. This reduction in her body's chemical energy is offset by a corresponding increase in the electrical potential energy $q\Delta V$. Not only that, but Alice feels the resistance stiffen as the charges get closer together and the repulsion strengthens. She has to do a little extra work, but this is all properly accounted for in the electrical potential energy.

But now suppose, b/3, that Betty decides to play a trick on Alice by tossing charge B far away just as Alice is getting ready to move charge A. We have already established that Alice can't feel charge B's motion instantaneously, so the electric forces must actually be propagated by an electric *field*. Of course this experiment is utterly impractical, but suppose for the sake of argument that the time it takes the change in the electric field to propagate across the diagram is long enough so that Alice can complete her motion before she feels the effect of B's disappearance. She is still getting stale information about B's position. As she moves A to the right, she



b / Fields carry energy.

feels a repulsion, because the field in her region of space is still the field caused by B in its *old* position. She has burned some chocolate cheesecake calories, and it appears that conservation of energy has been violated, because these calories can't be properly accounted for by any interaction with B, which is long gone.

If we hope to preserve the law of conservation of energy, then the only possible conclusion is that the electric field itself carries away the cheesecake energy. In fact, this example represents an impractical method of transmitting radio waves. Alice does work on charge A, and that energy goes into the radio waves. Even if B had never existed, the radio waves would still have carried energy, and Alice would still have had to do work in order to create them.

10.2.3 Fields must have transformation laws

In the foregoing discussion I've been guilty of making arguments that fields were "real." Sorry. In physics, and particularly in relativity, it's usually a waste of time worrying about whether some effect such as length contraction is "real" or only "seems that way." But thinking of fields as having an independent existence does lead to a useful guiding principle, which is that *fields must have transformation laws*. Suppose that at a certain location, observer \mathbf{o}_1 measures every possible field — electric, magnetic, bodice-ripper-sexual-attractiveness, and so on. (The gravitational field is not on the list, for the reasons discussed in section 5.2.) Observer \mathbf{o}_2 , passing by the same event but in a different state of motion, could carry out similar measurements. We're talking about measurements being carried out on a cubic inch of pure vacuum, but suppose that the answer to Peggy Lee's famous question is "Yes, that's all there is" — the only information there is to know about that empty parcel of nothingness is the (frame-dependent) value of the fields it contains. Then \mathbf{o}_1 ought to be able to predict the results of \mathbf{o}_2 's measurements. For if not, then what is the nature of the information that is hidden from \mathbf{o}_1 but revealed to \mathbf{o}_2 ? Presumably this would be something related to how the fields were produced by certain particles long ago and far away. For example, maybe \mathbf{o}_1 is at rest relative to a certain charge q that helped to create the fields, but \mathbf{o}_2 isn't, so \mathbf{o}_2 picks up q 's magnetic field, which is information unavailable to \mathbf{o}_1 — who thinks q was at rest, and therefore didn't *make* any magnetic field. This would contradict our "that's all there is" hypothesis.

To show the power of "that's all there is," consider example 1, p. 176, in which we found that boosting a solenoid along its own axis doesn't change its internal field. As a fact about solenoids, it's fairly obscure and useless. But if the fields must have transformation laws, then we've learned something much more general: a magnetic field *always* stays the same under a boost in the direction of the field.

10.3 Electromagnetic fields

10.3.1 The electric field

Section 10.1 showed that relativity requires magnetic *forces* to exist, and section 10.2.3 gave us a peek at what this implies about or how electric and magnetic *fields* transform. To understand this on a more general basis, let's explicitly list some assumptions about the electric field and see how they lead to the existence and properties of a magnetic field:

1. *Definition of the electric field:* In the frame of reference of an inertial observer \mathbf{o} , take some standard, charged test particle, release it at rest, and observe the force $\mathbf{F}_{\mathbf{o}}$ (section 4.5, p. 100) acting on the particle. (The timelike component of this force vanishes.) Then the electric field three-vector \mathbf{E} in frame \mathbf{o} is defined by $\mathbf{F}_{\mathbf{o}} = q\mathbf{E}$, where we fix our system of units by taking some arbitrary value for the charge q of the test particle.
2. *Definition of electric charge:* For charges other than the standard test charge, we take Gauss's law to be our definition of electric charge.
3. Charge is Lorentz invariant (p. 22).
4. Fields must have transformation laws (section 10.2.3).

Many times already in our study of relativity, we've followed the strategy of taking a Galilean vector and trying to redefine it as a four-dimensional vector in relativity. Let's try to do this with the electric field. Then we would have no other obvious thing to try than to change its definition to $\mathbf{F} = q\mathbf{E}$, where $\mathbf{F} = m\mathbf{a}$ is the relativistic force vector (section 4.5, p. 100), so that the electric field three-vector was just the spacelike part of \mathbf{E} . Because $\mathbf{a} \cdot \mathbf{v} = 0$ for a material particle, this would imply that \mathbf{E} was orthogonal to \mathbf{o} for any observer \mathbf{o} . But this is impossible, since then a spacetime displacement vector \mathbf{s} along the direction of \mathbf{E} would be a vector of simultaneity for all observers, and we know that this isn't possible in relativity.

10.3.2 The magnetic field

Our situation is very similar to the one encountered in section 9.1, p. 175, where we found that knowledge of the charge density in one frame was insufficient to tell us the charge density in other frames. There was missing information, which turned out to be the current density. The problems we've encountered in defining the transformation properties of the electric field suggest a similar "missing-information" situation, and it seems likely that the missing information is the magnetic field. How should we modify the assumptions on p. 218 to allow for the existence of a magnetic field

in addition to the electric one? What properties could this additional field have? How would we define or measure it?

One way of imagining a new type of field would be if, in addition to charge q , particles had some other characteristic, call it r , and there was then be some entirely separate field defined by their action on a particle with this “ r -ness.” But going down this road leads us to unrelated phenomena such as the the strong nuclear interaction.

10.3.3 The electromagnetic field tensor

The nature of the contradiction arrived at in section 10.3.1 is such that our additional field is closely linked to the electric one, and therefore we expect it to act on charge, not on r -ness. Without inventing something new like r -ness, the only other available property of the test particle is its state of motion, characterized by its velocity vector \mathbf{v} . Now the simplest rule we could imagine for determining the force on a test particle would be a linear one, which would look like matrix multiplication:

$$\mathbf{F} = q\mathcal{F}\mathbf{v}$$

or in index notation,

$$F^a = q\mathcal{F}^a_b v^b.$$

Although the form \mathcal{F}^a_b with one upper and one lower index occurs naturally in this expression, we’ll find it more convenient from now on to work with the upper-upper form \mathcal{F}^{ab} . \mathcal{F} would be 4×4 , so it would have 16 elements:

$$\begin{pmatrix} \mathcal{F}^{tt} & \mathcal{F}^{tx} & \mathcal{F}^{ty} & \mathcal{F}^{tz} \\ \mathcal{F}^{xt} & \mathcal{F}^{xx} & \mathcal{F}^{xy} & \mathcal{F}^{xz} \\ \mathcal{F}^{yt} & \mathcal{F}^{yx} & \mathcal{F}^{yy} & \mathcal{F}^{yz} \\ \mathcal{F}^{zt} & \mathcal{F}^{zx} & \mathcal{F}^{zy} & \mathcal{F}^{zz} \end{pmatrix}$$

Presumably these 16 numbers would encode the information about the electric field, as well as some additional information about the field or fields we were missing.

But these are not 16 numbers that we can choose freely and independently. For example, consider a charged particle that is instantaneously at rest in a certain observer’s frame, with $\mathbf{v} = (1, 0, 0, 0)$. (In this situation, the four-force equals the force measured by the observer.) The work done by a force is positive if the force is in the same direction as the motion, negative if in the opposite direction, and zero if there is no motion. Therefore the power $P = dW/dt$ in this example should be zero. Power is the timelike component of the force vector, which forces us to take $\mathcal{F}^{tt} = 0$.

More generally, consider the kinematical constraint $\mathbf{a} \cdot \mathbf{v} = 0$ (p. 64). When we require $\mathbf{a} \cdot \mathbf{v} = 0$ for *any* \mathbf{v} , not just this one, we end up with the constraint that \mathcal{F} must be antisymmetric, meaning

that when we transpose it, the result is another matrix that looks just like the original one, but with all the signs flipped:

$$\begin{pmatrix} 0 & \mathcal{F}^{tx} & \mathcal{F}^{ty} & \mathcal{F}^{tz} \\ -\mathcal{F}^{tx} & 0 & \mathcal{F}^{xy} & \mathcal{F}^{xz} \\ -\mathcal{F}^{ty} & -\mathcal{F}^{xy} & 0 & \mathcal{F}^{yz} \\ -\mathcal{F}^{tz} & -\mathcal{F}^{xz} & -\mathcal{F}^{yz} & 0 \end{pmatrix}$$

Each element equals minus the corresponding element across the main diagonal from it, and antisymmetry also requires that the main diagonal itself be zero. In terms of the concept of degrees of freedom introduced in section 3.5.3, p. 62, we are down to 6 degrees of freedom rather than 16.

We now relabel the elements of the matrix and follow up with a justification of the relabeling. The result is the following rank-2 tensor:

$$\mathcal{F}^{\mu\nu} = \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{pmatrix} \quad (1)$$

We'll call this the electromagnetic field tensor. The labeling of the left column simply expresses the definition of the electric field, which is expressed in terms of the velocity $\mathbf{v} = (1, 0, 0, 0)$ of a particle at rest. The top row then follows from antisymmetry. For an arbitrary velocity vector, writing out the matrix multiplication $F^\mu = q\mathcal{F}^\mu_\nu v^\nu$ results in expressions such as $F^x = \gamma q(E_x + u_y B_z - u_z B_y)$ (problem 3, p. 237). Taking into account the difference of a factor of γ between the four-force and the force measured by an observer, we end up with the familiar Lorentz force law,

$$\mathbf{F}_o = q(\mathbf{E} + \mathbf{u} \times \mathbf{B}),$$

where \mathbf{B} is the magnetic field. This is expressed in units where $c = 1$, so that the electric and magnetic field have the same units. In units with $c \neq 1$, the magnetic components of the electromagnetic field matrix should be multiplied by c .

Thus starting only from the assumptions on p. 218, we deduce that the electric field must be accompanied by a magnetic field.

Parity properties of \mathbf{E} and \mathbf{B}

Example 1

In example 6 on p. 181, we saw that under the parity transformation $(t, x, y, z) \rightarrow (t, -x, -y, -z)$, any rank-2 tensor expressed in Minkowski coordinates changes the signs of its components according to the same rule:

$$\begin{pmatrix} \text{no flip} & \text{flip} & \text{flip} & \text{flip} \\ \text{flip} & \text{no flip} & \text{no flip} & \text{no flip} \\ \text{flip} & \text{no flip} & \text{no flip} & \text{no flip} \\ \text{flip} & \text{no flip} & \text{no flip} & \text{no flip} \end{pmatrix}.$$

Since this holds for the electromagnetic field tensor \mathcal{F} , we find that under parity, $\mathbf{E} \rightarrow -\mathbf{E}$ and $\mathbf{B} \rightarrow \mathbf{B}$. For example, a capacitor seen in a mirror has its electric field pointing the opposite way, but there is no change in the magnetic field of a current loop, since the location of each current element is flipped to the other side of the loop, but its direction of flow is also reversed, so that the picture as a whole remains unchanged.

10.3.4 What about gravity?

A funny puzzle pops up if we go back and think about the assumptions on p. 218 that went into all this. Those assumptions were so general that it almost seems as though the only possible behavior for fields is the behavior of electric and magnetic fields. But other fields *do* behave differently. How did the assumptions fail in the case of gravity, for example? Gauss's law (assumption 2) certainly holds for gravity. But the source of gravitational fields isn't charge, it's mass-energy, and mass-energy isn't a Lorentz invariant, contrary to assumption 3. Furthermore, assumption 1 entailed that our field could be defined in terms of *forces* measured by an inertial observer, but for an inertial observer gravity doesn't exist (section 5.2).

10.4 Transformation of the fields

Since we have associated the components of the electric and magnetic fields with elements of a rank-2 tensor, the transformation law for these fields now follows from the general tensor transformation law for rank-2 tensors (p. 180). We first state the general rule, in a prettified form, and then give some concrete examples. Under a boost by a three-velocity \mathbf{v} , the electric and magnetic fields \mathbf{E} and \mathbf{B} transform to \mathbf{E}' and \mathbf{B}' according to these rules:

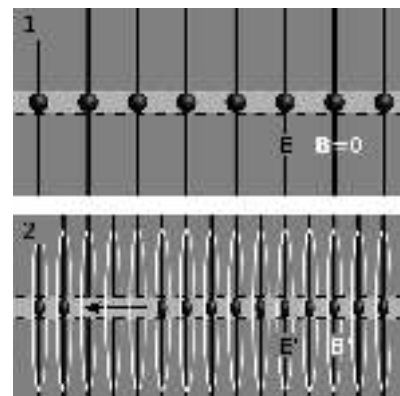
$$\begin{aligned} \mathbf{E}'_{\parallel} &= \mathbf{E}_{\parallel} & \mathbf{E}'_{\perp} &= \gamma(\mathbf{E}_{\perp} + \mathbf{v} \times \mathbf{B}) \\ \mathbf{B}'_{\parallel} &= \mathbf{B}_{\parallel} & \mathbf{B}'_{\perp} &= \gamma(\mathbf{B}_{\perp} - \mathbf{v} \times \mathbf{E}) \end{aligned}$$

A line of charge

Figure c/1 shows a line of charges. At a given nearby point, it creates an electric field \mathbf{E} that points outward, as measured by an observer \mathbf{o} who is at rest relative to the charges. This field is represented in the figure by its pattern of field lines, which start on the charges and radiate outward like the bristles of a bottle brush. Because the charges are at rest, the magnetic field is zero. (Finding the magnitude of the field at a certain distance is a straightforward application of Gauss's law.)

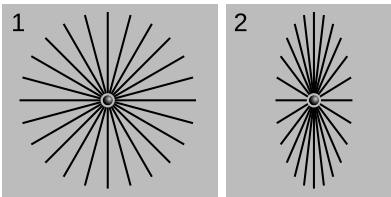
Now consider an observer \mathbf{o}' , figure c/2, moving at velocity \mathbf{v} to the right relative to \mathbf{o} . Without even worrying about how the field was created, we can transform the fields, at the point in space discussed previously, into the new frame. The result is $\mathbf{E}' = \gamma\mathbf{E}$

Example 2



c / Example 2.

and $\mathbf{B}'_{\perp} = -\gamma \mathbf{v} \times \mathbf{E}$. In this frame, the electric field is more intense, and there is also a magnetic field, whose pattern of white field lines forms circles lying in planes perpendicular to the line. If we do happen to know that the field was created by the line of charge, which is moving according to \mathbf{o}' , then we can explain these results as arising from two effects. First, the line of charge has been length-contracted. This causes the density of charge per unit length to increase by a factor of γ , with a proportional increase in the electric field. In the field-line description, we simply have more charges in the figure, so there are more field lines coming out of them. Second, the line of charge is moving to the left in this frame, so it forms an electric current, and this current is the cause of the magnetic field \mathbf{B}' .



d / Example 3.

A moving charge

Example 3

Figure d/1 shows the electric field lines of a charge, in the charge's rest frame K . In figure d/2 we see the same electric field, in a frame K' in which the charge is moving along the x axis, which points to the right, at 90% of c . (In this frame there is also a magnetic field, which is not shown.) This electric field, which is time-varying, is shown as a snapshot in a hyperplane of simultaneity $t' = 0$ of K' . Surprisingly, these field lines all point toward the charge's *present* position in K' .

Disturbances in the electromagnetic field propagate at c , not instantaneously, so one might have expected the field at a certain location P in this figure to point toward a location at a distance r that the charge had occupied at an earlier time $t' = -r/c$. This would have produced a set of curved field lines reminiscent of the wake of a boat. To see that this is not possible, consider the point $(0, 0, h, 0)$ in the Minkowski coordinates of K , i.e., a point on the y axis. After a Lorentz transformation along x , the coordinates of this point in K' are still $(0, 0, h, 0)$, so in K' as well it lies on a line that passes transversely through the present position of the charge. Since this point has $E_x = 0$ and $\mathbf{B} = 0$ in K , application of the transformation laws shows that $E'_x = 0$ as well, so that the field points toward the charge's present position, not its past position.

A similar but more complicated calculation shows that the field at intermediate angles is also in the instantaneously radial direction. Rather than filling in the details, we note that this makes sense because the Poynting vector $\mathbf{E} \times \mathbf{B}$ then has no radial component, which is as expected because energy should be transported forward but not radiated outward.

One might worry that this would indicate that the information about the charge's position was propagating instantaneously, contradicting relativity. But this is a charge that has always been in its

current state of motion and always will be. If the charge's motion had been disturbed by some external force at a time later than $t' = -r/c$, the field lines in K would still be pointing toward the location that the charge had previously occupied while at rest, and the field in K' would be pointing toward its linearly extrapolated position.

A field behaving like a stick

Example 4

Figure d/2 appears identical to a copy of figure d/1 that has been Lorentz contracted by $1/\gamma$, and we can verify from the transformation laws for the fields that this is correct. Since these transformation laws apply regardless of how the fields were produced, we have a general rule, which is that if a field is purely electric in one frame, then its direction transforms to another frame in the same way as the direction of a stick, when we transform out of the stick's rest frame. (See problem 3, p. 51.)

It is *not* true in general that electric field lines can simply be carried over from one frame to another as if we were Lorentz-contracting a rat's nest built out of wire. This property holds only when the original frame is of a very special kind: a frame in which the field is purely electric. (We can always find such a frame if $E^2 > B^2$; see section 10.5.) As a counterexample to the notion that it applies more generally, consider the case in which a field is purely *magnetic* in a certain frame. Then the electric field lines do not even exist in the original frame, but do exist in the new one.

Coming back to the case where the original field is purely electric, so that the stick-like behavior does hold, it is not immediately obvious why there should be this strange correspondence between sticks and field lines. The methods used in problem 3 do not seem to have much in common with the ones we have used to determine how the electric field behaves. But the following physical argument shows that there is a simple reason for the identical behavior.

Consider a stick with charges $+q$ and $-q$ fixed at the ends. The stick is nonrotating and moving inertially. In the stick's rest frame K , there is a field line originating from $+q$ and terminating on $-q$ which coincides with the stick. Now consider frame K' moving in some direction relative to the stick. As discussed in example 3, the field due to each charge points toward or away from its present instantaneous position in K' as well as K . Therefore each field, at the stick, is parallel to the stick, and we again have a field line in K' that coincides with the stick. Since the transformation of the field is independent of how the field was created, this holds for any field that is purely electric in the original frame.

10.5 Invariants

We've seen cases before in which an invariant can be formed from a rank-1 tensor. The square of the proper time corresponding to a timelike spacetime displacement \mathbf{r} is $\mathbf{r} \cdot \mathbf{r}$ or, in the index notation introduced in section ??, $r^a r_a$. From the momentum tensor we can construct the square of the mass $p^a p_a$.

There are good reasons to believe that something similar can be done with the electromagnetic field tensor, since electromagnetic fields have certain properties that are preserved when we switch frames. Specifically, an electromagnetic wave consists of electric and magnetic fields that are equal in magnitude and perpendicular to one another. An electromagnetic wave that is a valid solution to Maxwell's equations in one frame should also be a valid wave in another frame. It can be shown that the following two quantities are invariants:

$$P = B^2 - E^2$$

and

$$Q = \mathbf{E} \cdot \mathbf{B}.$$

The fact that these are written as vector dot products of three-vectors shows that they are invariant under rotation, but we also want to show that they are relativistic scalars, i.e., invariant under boosts as well. To prove this, we can write them both in tensor notation. The first invariant can be expressed as $P = \frac{1}{2} \mathcal{F}^{ab} \mathcal{F}_{ab}$, while the second equals $Q = \frac{1}{4} \epsilon^{abcd} \mathcal{F}_{ab} \mathcal{F}_{cd}$, where $\epsilon^{\kappa\lambda\mu\nu}$ is the Levi-Civita tensor.

A field for which both $P = Q = 0$ is called a null field. An electromagnetic plane wave is a null field, and although this is easily verifiable from the definitions of P and Q , there is a deeper reason why this should be true, and this reason applies not just to electromagnetic waves but to other types of waves, such as gravitational waves. Consider any relativistic scalar s that is a continuous function of the electromagnetic field tensor \mathcal{F} , i.e., a continuous function of \mathcal{F} 's components. We want s to vanish when $\mathcal{F} = 0$. Given an electromagnetic plane wave, we can do a Lorentz boost parallel to the wave's direction of propagation. Under such a boost the wave suffers a Doppler shift in its wavelength and frequency, but in addition to that, the transformation equations on p. 221 imply that the intensity of the fields is reduced at any given point. Thus in the limit of an indefinite process of acceleration, $\mathcal{F} \rightarrow 0$, and therefore $s \rightarrow 0$ as well. But since s is a scalar, its value is independent of our frame of reference, and so it must be zero in all frames.

P and Q are a complete set of invariants for the electromagnetic field, meaning that the only other electromagnetic invariants are those that either can be determined from P and Q or depend on the derivatives of the fields, not just their values. To see that P and Q are complete in this sense, we can break the possibilities down into cases, according to whether P and Q are zero or nonzero, positive or negative. As a representative example, consider the case where $P < 0$ and $Q > 0$. First we rotate our frame of reference so that \mathbf{E} is along the x axis, and \mathbf{B} lies in the x - y plane. Next we do a boost along the z axis in order to eliminate the y component of \mathbf{B} ; the field transformation equations on p. 221 make this possible because $|\mathbf{E}| > |\mathbf{B}|$. The result is that we have found a frame of reference in which \mathbf{E} and \mathbf{B} both lie along the positive x axis. The only frame-independent information that there is to know is the information available in this frame, and that consists of only two positive real numbers, E_x and B_x , which can be determined from the values of P and Q .

A static null field

Example 5

Although an electromagnetic plane wave is a null field, the converse is not true. For example, we can create a static null field out of a static, uniform electric field and a static, uniform magnetic field, with the two fields perpendicular to one another.

Another invariant?

Example 6

Let Π be the squared magnitude of the Poynting vector, $\Pi = (\mathbf{E} \times \mathbf{B}) \cdot (\mathbf{E} \times \mathbf{B})$. Since Π can be expressed in terms of dot products and scalar products, it is guaranteed to be invariant under rotations. However, it is not a relativistic invariant. For example, if we do a Lorentz boost parallel to the direction of an electromagnetic wave, the intensity of the wave changes, and so does Π .

A non-null invariant for electromagnetic waves?

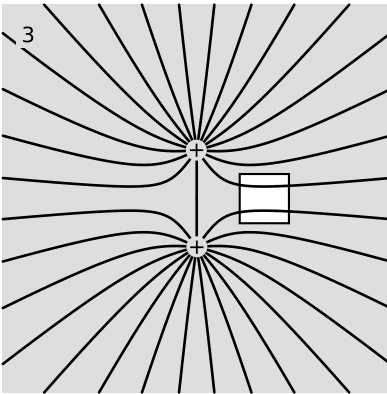
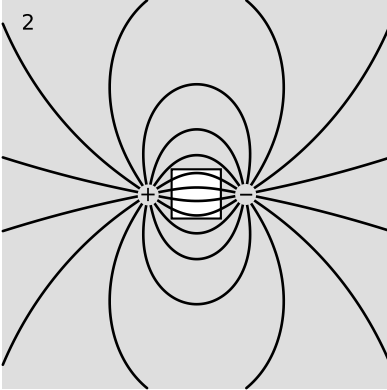
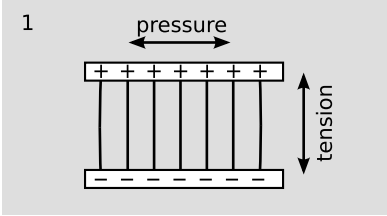
Example 7

The quantity $Q^{-1} = 1/(\mathbf{E} \cdot \mathbf{B})$ is clearly an invariant, and it doesn't vanish for an electromagnetic plane wave — in fact, it is infinite for a plane wave. Does this contradict our proof that any invariant must vanish for a plane wave? No, because we only proved this in the case where the invariant is defined as a continuous function of \mathcal{F} . Our function Q^{-1} is a discontinuous function of \mathcal{F} when $\mathcal{F} = 0$. Such discontinuous invariants tend not to be very interesting. For suppose we try to measure Q^{-1} , and the thing we're measuring happens to be an electromagnetic wave. Our measurements of the fields will probably be statistically consistent with zero, and therefore the error bars on our measurement of Q^{-1} will likely be infinitely large.

10.6 Stress-energy tensor of the electromagnetic field

The electromagnetic field has a stress-energy tensor associated with it. From our study of electromagnetism we know that the electromagnetic field has energy density $U = (E^2 + B^2)/8\pi k$ and momentum density $\mathbf{S} = (\mathbf{E} \times \mathbf{B})/4\pi k$ (in units where $c = 1$, with k being the Coulomb constant). This fixes the components of the stress-energy tensor of the form $T^{t\dots}$ and $T^{\dots t}$, i.e., the top row and left column, to look like this:

$$T^{\mu\nu} = \begin{pmatrix} U & S_x & S_y & S_z \\ S_x & & & \\ S_y & & & \\ S_z & & & \end{pmatrix}.$$



e / Pressure and tension in electrostatic fields.

The following argument tells us something about what to expect for the components T^{xx} , T^{yy} , and T^{zz} , which are interpreted as pressures or tensions, depending on their signs. In figure e/1, the capacitor plates want to collapse against each other in the vertical (y) direction, but at the same time the internal repulsions within each plate make that plate want to expand in the x direction. If the capacitor is built out of materials that hold their shape, then the electromagnetic tension in $T^{yy} < 0$ is counteracted by pressure $T^{yy} > 0$ in the materials, while the electromagnetic pressure $T^{xx} > 0$ is canceled by the materials' tension $T^{xx} < 0$. We got these results for a particular physical situation, but relativity requires that the stress-energy be defined at every point based on the fields at that point, so our conclusions must hold generally. In e/2 and e/3, white boxes have been drawn in regions where the total field is strong and the fields are strongly interacting. In 2, there is tension in the x direction and pressure in y ; the tension can be thought of as contributing to the attraction between the opposite charges. In 3, there is also x tension and y pressure; the pressure contributes to the like charges' repulsion.

To make this more quantitative, consider the discontinuity in E_y at the upper plate in figure e/1. The field abruptly switches from 0 on the outside to some value E between the plates. By Gauss's law, the charge per unit area on the plate must be $\sigma = E/4\pi k$. The average field experienced by the charge in the plate is $\bar{E} = (0 + E)/2 = E/2$, so the force per unit area, i.e., the tension in the field, is $\sigma\bar{E} = E^2/8\pi k$. Thus we expect $T^{yy} = -E^2/8\pi k$ if E is along the y axis.

For the reader who wants the full derivation of the remaining nine components of the tensor, we now give an argument that makes use of the following list of its properties. Other readers can skip ahead to where the full tensor is presented.

1. T is symmetric, $T^{\mu\nu} = T^{\nu\mu}$.
2. The components must be second-order in the fields, e.g., we can have terms like $E_x B_z$, but not $E_x^3 B_z^7$ or $E_x B_z B_y$. This is because Maxwell's equations are linear, and when a wave equation is perfectly linear, the corresponding energy expression is second-order in the amplitude of the wave.
3. T has the parity properties described in example 6 on p. 181.
4. The electric and magnetic fields are treated symmetrically in Maxwell's equations, so they should be treated symmetrically in the stress-energy tensor. E.g., we could have a term like $7E_x^2 + 7B_z^2$, but not $7E_x^2 + 6B_z^2$.
5. On p. 187 of section 9.2.8, we saw that the trace energy condition $T^a_a \geq 0$ is satisfied by a cloud of dust if and only if the dust's mass-energy is not transported at a speed greater than c . In section 4.1, we saw that all ultrarelativistic particles have the same mechanical properties. Since a cloud of dust, in the limit where its speed approaches c , is on the edge of the bound set by the trace energy condition, $T^a_a \rightarrow 0$, we expect that the electromagnetic field, in which disturbances propagate at c , should also exactly saturate the trace energy condition, so that $T^a_a = 0$.
6. The stress-energy tensor should behave properly under rotations, which basically means that x , y , and z should be treated symmetrically.
7. An electromagnetic plane wave propagating in the x direction should not exert any pressure in the y or z directions.
8. If the field obeys Maxwell's equations, then the energy-conservation condition $\partial T^{ab}/\partial x^a = 0$ should hold.

These facts are enough to completely determine the form of the remaining nine components of the stress-energy tensor. Property 3 requires that all of these components be even under parity. Since electric fields flip under parity but magnetic fields don't (example 1, p. 220), these components can only have terms like $E_i E_j$ and $B_i B_j$, not mixed terms like $E_i B_j$. Taking into account properties 4 and 6, we find that the diagonal terms must look like

$$4\pi k T^{xx} = a(E_x^2 + B_x^2) + b(E^2 + B^2),$$

and the off-diagonal ones

$$4\pi k T^{xy} = c(E_x E_y + B_x B_y).$$

Property 5 gives $1/2 - a - 3b = 0$ and 7 gives $b = -a/2$, so we have $a = -1$ and $b = 1/2$. The determination of $c = -1$ is left as an exercise, problem 4 on p. 237.

We have now established the complete expression for the stress-energy tensor of the electromagnetic field, which is

$$T^{\mu\nu} = \begin{pmatrix} U & S_x & S_y & S_z \\ S_x & -\sigma_{xx} & -\sigma_{xy} & -\sigma_{xz} \\ S_y & -\sigma_{yx} & -\sigma_{yy} & -\sigma_{yz} \\ S_z & -\sigma_{zx} & -\sigma_{zy} & -\sigma_{zz} \end{pmatrix},$$

where

$$U = \frac{1}{8\pi k}(E^2 + B^2),$$

$$\mathbf{S} = \frac{1}{4\pi k}\mathbf{E} \times \mathbf{B},$$

and σ , known as the Maxwell stress tensor, is given by

$$-\sigma_{\mu\nu} = \frac{1}{4\pi k} \times \begin{cases} -E_\mu E_\nu - B_\mu B_\nu & \text{if } \mu \neq \nu \\ -E_\mu E_\nu - B_\mu B_\nu + \frac{1}{2}(E^2 + B^2) & \text{if } \mu = \nu \end{cases}$$

All of this can be expressed more compactly and in a coordinate-independent way as

$$T^{ab} = \frac{1}{4\pi k} \left(\mathcal{F}^{ac} \mathcal{F}^b_c + \frac{1}{4} o^d o_d g^{ab} \mathcal{F}_{ef} \mathcal{F}^{ef} \right), \quad (2)$$

where \mathbf{o} is a future-directed velocity vector, so that $o^d o_d = +1$ for the signature $+- --$ used in this book, and -1 if the signature is $-+++$.

Stress-energy tensor of a plane wave

Example 8

Let an electromagnetic plane wave (not necessarily sinusoidal) propagate along the x axis, with its polarization such that \mathbf{E} is in the y direction and \mathbf{B} on the z axis, and $|\mathbf{E}| = |\mathbf{B}| = A$. Then we have the following for the stress-energy tensor.

$$T^{\mu\nu} = \frac{A^2}{4\pi k} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

The T^{tt} component tells us that the wave has a certain energy density. Because the wave is massless, we have $E^2 - p^2 = m^2 = 0$, so the momentum density is the same as the energy density, and T^{tx} is the same as T^{tt} . If this wave strikes a surface in the yz plane, the momentum the surface absorbs from the wave will be felt as a pressure, represented by T^{xx} .

In example 5 on p. 181, we saw that a cloud of dust, viewed in a frame moving at velocity v relative to the dust's rest frame, had the following stress-energy tensor.

$$T^{\mu\nu} = \begin{pmatrix} \gamma^2 \rho & \gamma^2 v \rho & 0 & 0 \\ \gamma^2 v \rho & \gamma^2 v^2 \rho & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

In the ultrarelativistic limit $v \rightarrow 1$, this becomes

$$T^{\mu\nu} = (\text{energy density}) \times \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

which is exactly the same as the result for our electromagnetic wave. This illustrates the fact discussed in section 4.1 that all ultrarelativistic particles have the same mechanical properties.

Mass of a capacitor

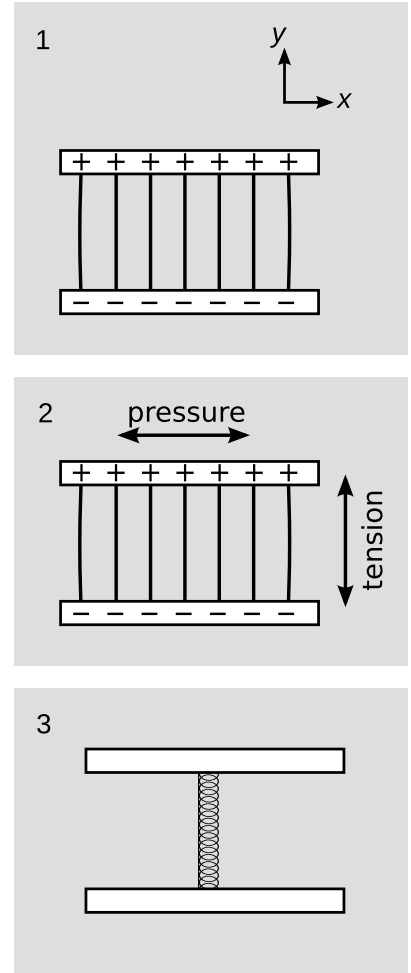
Example 9

Consider the mass of a charged parallel-plate capacitor, figure f/1, first in its rest frame and then in a frame boosted in the direction parallel to the field (perpendicular to the plates). If we're not careful, we run into the following paradox. Under a boost, an electric field parallel to the boost remains unchanged. Therefore in the boosted frame, we have exactly the same field strength, but filling a volume that has been decreased by length contraction. Therefore the mass-energy of the capacitor is *greatest* in its own rest frame, which is absurd and would contradict our proof in section 9.3.4 that the energy-momentum of an isolated system transforms as a four-vector.

The resolution of the paradox comes from recognizing that we assumed the capacitor to be in static equilibrium, but we ignored the stress-energy of whatever mechanical supports were maintaining this equilibrium. If we consider only the stress-energy $T_{(em)}$ of the electromagnetic field, then we have $T_{(em)}^{tt} = (1/8\pi k)E^2$ (energy density) and $T_{(em)}^{yy} = -(1/8\pi k)E^2$ (tension in the y direction, parallel to the field), figure f/2. It's easy to see that this has a nonvanishing divergence, since $\partial_y T_{(em)}^{yy} \neq 0$ at the plates, and there are no other terms in the stress-energy tensor that could compensate for this.

There is nothing surprising here; only the *total* stress-energy tensor T has to be divergenceless, not $T_{(em)}$. It would violate the laws of physics if the capacitor were to remain in equilibrium like this without some force to counter the electromagnetic tension. Let's say that this force is provided by a spring, as in figure f/3. The spring has its own contribution $T_{(s)}$ to the stress-energy. For convenience, let's imagine making the spring filled in (rather than a hollow cylinder) and fattening it up so that it fills the entire interior volume of the capacitor. Then to achieve static equilibrium in the rest frame, we need the pressure in the spring to cancel out the pressure in the electric field. We therefore have $T^{yy} = 0$ for the total stress-energy tensor.

If we now apply the tensor transformation law to the stress-energy tensor, we find that the stress-energy tensor in the boosted frame contains a mass-energy density $T^{t't'}$ that depends only on T^{tt}



f / Example 9.

and T^{yy} . (There also has to be an xx component to keep the plates from exploding laterally, but that doesn't enter here.) But we have $T^{yy} = 0$, so the problem is exactly the same as transforming a lump of nonrelativistic matter, and we know that that calculation comes out OK. For an explicit demonstration that this still works out if we drop the simplifying assumption that the spring fills the entire interior volume of the capacitor, see Rindler and Denur, "A simple relativistic paradox about electrostatic energy," *Am J Phys* 56 (1988) 9.

10.7 Maxwell's equations

10.7.1 Statement and interpretation

In this book I assume that you've had the usual physics background acquired in a freshman survey course, which includes an initial, probably frightening, encounter with Maxwell's equations in integral form. In units with $c = 1$, Maxwell's equations are:

$$\Phi_E = 4\pi kq \quad (3a)$$

$$\Phi_B = 0 \quad (3b)$$

$$\oint \mathbf{E} \cdot d\boldsymbol{\ell} = -\frac{\partial \Phi_B}{\partial t} \quad (3c)$$

$$\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \frac{\partial \Phi_E}{\partial t} + 4\pi kI \quad (3d)$$

where

$$\Phi_E = \int \mathbf{E} \cdot d\mathbf{a} \quad \text{and} \quad (4)$$

$$\Phi_B = \int \mathbf{B} \cdot d\mathbf{a}. \quad (5)$$

Equations (3a) and (3b) refer to a closed surface and the charge q contained inside that surface. Equation (3a), Gauss's law, says that charges are the sources of the electric field, while (3b) says that magnetic "charges" don't exist. Equations (3c) and (3d) refer to a surface like a potato chip, which has an edge or boundary, and the current I passing through that surface, with the line integrals in being evaluated along that boundary. The right-hand side of (3c) says that a changing magnetic field produces a curly electric field, as in a generator or a transformer. The I term in (3d) says that currents create magnetic fields that curl around them. The $\partial \Phi_E / \partial t$ term, which says that changing electric fields create magnetic fields, is necessary so that the equations produce consistent

results regardless of the surfaces chosen, and is also part of the apparatus responsible for the existence of electromagnetic waves, in which the changing \mathbf{E} field produces the \mathbf{B} , and the changing \mathbf{B} makes the \mathbf{E} .

Equations (3a) and (3b) have no time-dependence. They function as constraints on the possible field patterns. Equations (3c) and (3d) are dynamical laws that predict how an initial field pattern will evolve over time. It can be shown that if (3a) and (3b) are satisfied initially, then (3c) and (3d) ensure that they will continue to be satisfied later. Because the dynamical laws consist of two vector equations, they provide a total of 6 constraints, which are the number needed in order to predict the behavior of the 6 fields E_x , E_y , E_z , B_x , B_y , and B_z .

10.7.2 Experimental support

Before Einstein's 1905 paper on relativity, the known laws of physics were Newton's laws and Maxwell's equations (3a)-(3d). Experiments such as example 4 on p. 85 show that Newton's laws are only low-velocity approximations. Maxwell's equations are *not* low-velocity approximations; for example, in section 1.3.1 we noted the evidence that atoms are electrically neutral, in agreement with Gauss's law, (3a), to one part in 10^{21} , even though the electrons in atoms typically have velocities on the order of 1-10% of c .

10.7.3 Incompatibility with Galilean spacetime

Maxwell's equations are not compatible with the Galilean description of spacetime (section 1.1.2, p. 13). If we assume that equations (3) hold in some frame \mathbf{o} , and then apply a Galilean boost, transforming the coordinates (t, x, y, z) to $(t', x', y', z') = (t, x - vt, y, z)$, we find that in frame \mathbf{o}' the equations have a different and more complicated form that cannot be simplified so as to look like the form they had in \mathbf{o} . Rather than writing out the resulting horrible mess and verifying that it can't be cast back into the simpler form, an easier way to prove this is to note that there are solutions to the equations in \mathbf{o} that are not solutions after a Galilean boost into \mathbf{o}' , if we try to keep the equations in the same form. For example, if a light wave propagates in the x direction at speed c in \mathbf{o} , then after a boost with $v = c$, we would have a light wave in frame \mathbf{o}' that was standing still. (This is Einstein's thought experiment of riding alongside a light wave on a motorcycle, p. 13.) Such a wave would violate (3c), since the left-hand side would be nonzero for a surface lying in the xy plane, but the time derivative on the right-hand side would be zero.

10.7.4 Not manifestly relativistic in their original form

Since Maxwell's equations are not low-velocity approximations and are incompatible with Galilean relativity, we expect with the benefit of historical hindsight that they *are* compatible with the

relativistic picture of spacetime. But when they are expressed in the form (3), they have two features, either one of which seems enough to make them completely *incompatible* with relativity:

- (i) They appear to describe instantaneous action at a distance. For example, Gauss's law, $\Phi_E = 4\pi kq$, relates the electric field in one place (on the closed surface) to the electric charge somewhere else (inside the surface). This *nonlocal* structure smells wrong relativistically, for the reasons discussed in section 10.2.
- (ii) They appear to treat time and space asymmetrically.

What's really happening here is that equations (3) are like a version of *Hamlet* written in crayon on a long strip of toilet paper. They are completely relativistic, but have been written in a form that hides that fact.

The problem of nonlocality, i, can be shown to be a non-issue because Maxwell's equations can be reworked into a form in which they are purely local. The idea is shown in figure g. The magnetic field lines all form closed loops, except for one of them, which begins at a point in space and extends off to infinity. Drawing the large box, 1, we find that $\Phi_{B,1}$, the flux of the magnetic field through the box, is not zero, because a line leaves the box but none come in. But the same discrepancy could have been detected with the smaller box 2, or in fact with an arbitrarily small box containing the source of the field line. In other words, the equation $\Phi_B = 0$ is nonlocal, but if it is to hold for any surface, then it must also hold locally, in the limit of an arbitrarily small surface. This purely *local* law of physics can be expressed using the three-dimensional version of the divergence, introduced on p. 178:

$$\frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} = 0$$

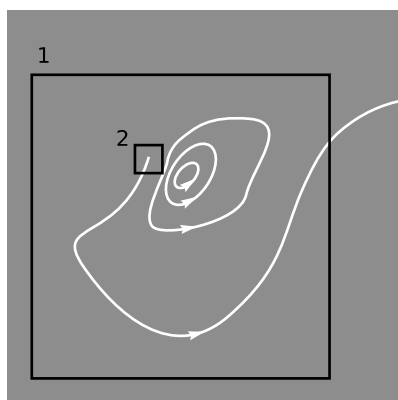
Of the four Maxwell's equations, both equation (3a) and (3b) can be reexpressed in this way. This book neither presents the full machinery of vector calculus nor assumes previous knowledge of it, but a similar limiting procedure can also be applied to equations (3c) and (3d), using an operator called the curl.

The following example is one in which both problem i and problem ii turn out not to be problems.

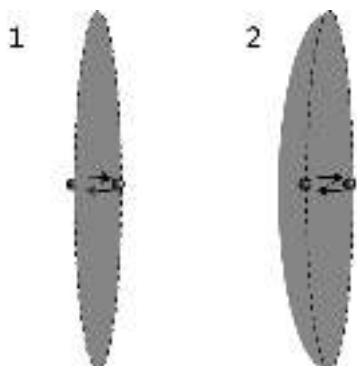
Jumping through a hoop

Example 10

Here is an example in which the non-obvious features of Maxwell's equations prevent the antirelativistic meltdown projected in i. In figure h/1, an electron jumps back and forth through an imaginary circular hoop, across which we construct an imaginary flat surface. Every time the electron pierces the surface, it makes a



g / A magnetic field that violates $\Phi_B = 0$.



h / 1. An electron jumps through a hoop. 2. An alternative surface spanning the hoop.

momentary spike in the current I , which appears in (3d),

$$\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \frac{\partial \Phi_E}{\partial t} + 4\pi kI.$$

We might expect that this would cause the field \mathbf{B} detected on the edge of the disk to show similar spikes at the same times. But “same times” implies some notion of simultaneity, and this would be incompatible with relativity, since the t coordinate being referred to here is just one observer’s notion of time. Furthermore, it would seem that information was being transmitted instantaneously from the center of the disk to its edge, which violates relativity (p. 17).

Stranger still, we can produce an apparent paradox without even appealing to relativity. Instead of the flat surface in h/1, we can pick a dish-shaped one, h/2, with a deep enough curve so that the electron never crosses it. The current I is always zero according to this surface, so that no field \mathbf{B} would be produced at the rim at all.

The resolution of all these difficulties lies in the term $\partial\Phi_E/\partial t$, which we’ve ignored. With surface 1, the electron crosses the surface in time δt , causing a current $I = e/\delta t$ but also causing a change in the flux from $\Phi_E \approx 2\pi ke$ to $\Phi_E \approx -2\pi ke$. The result is that the right-hand side of the equation is nearly zero. With surface 2, $I = 0$ and $\partial\Phi_E/\partial t \approx 0$, so the right-hand side is again nearly zero.

When the approximations used above are eliminated, Maxwell’s equations do predict a nonvanishing field, which is the expected electromagnetic wave propagating away from the electron at the proper speed c .

10.7.5 Lorentz invariance

Example 10 might seem like a “just-so story,” but the apparently miraculous resolution is not a coincidence. It happens because Maxwell’s equations are in fact invariant under a Lorentz transformation, even though that isn’t obvious when they’re written in the form (3a)-(3d). There are various ways of showing this:

- Einstein did it by brute force in his 1905 paper on relativity, by transforming the coordinates through a Lorentz transformation and the fields as in section 10.4.
- Maxwell’s equations are basically wave equations. (They have both wave solutions and static solutions.) We can verify that when we start with a sinusoidal plane wave in one frame, then transform into another frame, the result is again a valid sine-wave solution, having been subjected to a Doppler shift (section 3.2) and aberration (section 6.5). This requires checking

that the wave is still purely transverse, but that follows easily from examining the invariants described in section 10.5. By a celebrated mathematical result called Fourier's theorem, *any* well-behaved wave can be written as a sum of sine waves, and therefore any wave solution of Maxwell's equations in one frame is also a solution in every other frame.

- Maxwell's equations can be rewritten in terms of tensors, obeying all the grammatical rules of index gymnastics. If they can be written in this form, they are automatically Lorentz invariant.

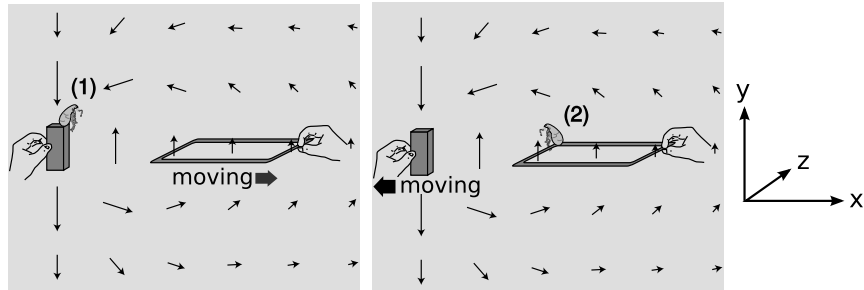
The last approach is the most general and elegant, so we'll provide a brief sketch of how it works. Equation (3a) has 4π times the charge on the right, while (3d) has 4π times the current. These both relate to the current four-vector \mathbf{J} , so clearly we need to combine them somehow into a single equation with \mathbf{J} on the right. Since the local form of equation (3a) involves the three-dimensional divergence, which contains first derivatives, the left-hand side of this combined equation should have a first derivative in it. Given the grammatical rules of tensors and index gymnastics, we don't have many possible ways to accomplish this. The only obvious thing to try is

$$\frac{\partial \mathcal{F}^{\mu\nu}}{\partial x^\nu} = 4\pi k J^\mu. \quad (6)$$

Writing this out for μ being the time coordinate, we get a relation that equates the divergence of \mathbf{E} to 4π times the charge density; this is the local equivalent of (3a). If you've taken vector calculus and know about the curl operator and Stokes' theorem, then you can verify that for μ referring to x , y , and z , we recover the local form of (3d). The tensorial way of expressing (3b) and (3c) turns out to be

$$\frac{\partial \mathcal{F}^{\mu\nu}}{\partial x^\lambda} + \frac{\partial \mathcal{F}^{\nu\lambda}}{\partial x^\mu} + \frac{\partial \mathcal{F}^{\lambda\mu}}{\partial x^\nu} = 0. \quad (7)$$

i / Example 11.



A generator

Example 11

Figure i shows a crude, impractical generator, depicted in two frames of reference.

Flea 1 is sitting on top of the bar magnet, which creates the magnetic field pattern shown with the arrows. To her, the bar magnet is obviously at rest, and this magnetic field pattern is static. As the square wire loop is dragged away from her and the magnet, its protons experience a force in the $-z$ direction, as determined by the Lorentz force law $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$. The electrons, which are negatively charged, feel a force in the $+z$ direction. The conduction electrons are free to move, but the protons aren't. In the front and back sides of the loop, this force is perpendicular to the wire. In the right and left sides, however, the electrons are free to respond to the force. Since the magnetic field is weaker on the right side, current circulates around the loop.

Flea 2 is sitting on the loop, which she considers to be at rest. In her frame of reference, it's the bar magnet that is moving. Like flea 1, she observes a current circulating around the loop, but unlike flea 1, she cannot use magnetic forces to explain this current. As far as she is concerned, the electrons were initially at rest. Magnetic forces are forces between moving charges and other moving charges, so a magnetic field can never accelerate a charged particle starting from rest. A force that accelerates a charge from rest can only be an *electric* force, so she is forced to conclude that there is an electric field in her region of space. This field drives electrons around and around in circles — it is a curly field. What reason can flea 2 offer for the existence of this electric field pattern? Well, she's been noticing that the magnetic field in her region of space has been changing, possibly because that bar magnet over there has been getting farther away. She observes that a changing magnetic field creates a curly electric field. Thus the $\partial\Phi_B/\partial t$ term in equation (3c) is not optional; it is required to exist if Maxwell's equations are to be equally valid in all frames.

Einstein opens his 1905 paper on relativity¹ begins with this sentence: "It is known that Maxwell's electrodynamics—as usually understood at the present time—when applied to moving bodies, leads to asymmetries which do not appear to be inherent in the phenomena." He then gives essentially this example. Although the observers in frames 1 and 2 agree on all physical measurements, their explanations of the physical mechanisms, couched in the language of Maxwell's equations in the form (3), are completely different. In relativistic language, flea 2's explanation can be written in terms of equation (7), in the case where the indices are x , z , and t :

$$\frac{\partial \mathcal{F}^{xz}}{\partial t} + \frac{\partial \mathcal{F}^{zt}}{\partial x} + \frac{\partial \mathcal{F}^{tx}}{\partial z} = 0,$$

¹"Zur Elektrodynamik bewegter Körper," *Annalen der Physik*. 17 (1905) 891. Translation by Perrett and Jeffery

which is the same as

$$\frac{\partial B_y}{\partial t} + \frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z} = 0.$$

Because the first term is negative, the second term must be positive. Since equations (6) and (7) are written in terms of tensors, obeying the grammatical rules of index gymnastics, we are *guaranteed* that they give consistent predictions in all frames of reference.

Conservation of charge and energy-momentum *Example 12*
Solving equation (6) for the current vector, we have

$$J^\mu = \frac{1}{4\pi k} \frac{\partial \mathcal{F}^{\mu\nu}}{\partial x^\nu}.$$

Conservation of charge (section 9.1.2, p. 178) can be expressed as

$$\frac{\partial J^\mu}{\partial x^\mu} = 0.$$

If we substitute the first equation into the second, we obtain

$$\frac{\partial}{\partial x^\mu} \left(\frac{1}{4\pi k} \frac{\partial \mathcal{F}^{\mu\nu}}{\partial x^\nu} \right) = 0$$

or

$$\frac{\partial^2 \mathcal{F}^{\mu\nu}}{\partial x^\mu \partial x^\nu} = 0,$$

with a sum over both μ and ν . But this equation is automatically satisfied because \mathcal{F} is antisymmetric, so for every combination of indices μ and ν , the term involving $\mathcal{F}^{\mu\nu}$ is canceled by one containing $\mathcal{F}^{\nu\mu} = -\mathcal{F}^{\mu\nu}$. Thus conservation of charge does not have to be added as a supplementary condition in addition to Maxwell's equations; it is automatically implied by Maxwell's equations.

Using equation (2) on p. 228, one can also prove that Maxwell's equations imply conservation of energy-momentum.

Problems

- 1** (a) A parallel-plate capacitor has charge per unit area $\pm\sigma$ on its two plates. Use Gauss's law to find the field between the plates. (b) In the style of example 2 on p. 221, transform the field to a frame moving perpendicular to the plates, and verify that the result makes sense in terms of the sources that are present. (c) Repeat the analysis for a frame moving parallel to the plates.

2 We've seen examples such as figure a on p. 215 in which a purely magnetic field in one frame becomes a mixture of magnetic and electric fields in another, and also cases like example 2 on p. 221 in which a purely electric field transforms to a mixture. Can we have a case in which a purely electric field in one frame transforms to a purely magnetic one in another? The easy way to do this problem is by using invariants.

3 (a) Starting from equation (1) on p. 220 for $\mathcal{F}^{\mu\nu}$, lower an index to find $\mathcal{F}^\mu{}_\nu$. Assume Minkowski coordinates and metric signature $+- - -$.

(a) Let $\mathbf{v} = \gamma(1, u_x, u_y, u_z)$, where (u_x, u_y, u_z) is the velocity three-vector. Write out the matrix multiplication $F^\mu = q\mathcal{F}^\mu{}_\nu v^\nu$, and show as claimed on p. 220 that the result is the Lorentz force law.

4 On p. 226 I presented a list of properties of the electromagnetic stress tensor, followed by an argument in which the tensor is constructed with three unknown constants a , b , and c , to be determined from those properties. The values of a and b are derived in the text, and the purpose of this problem is to finish up by proving that $c = -1$. The idea is to take the field of a point charge, which we know satisfies Maxwell's equations, and then apply property 8, which requires that the energy-conservation condition $\partial T^{ab}/\partial x^a = 0$ hold. This works out nicely if you apply this property to the x column of T , at a point that lies in the positive x direction relative to the charge.

5 Show that the number of independent conditions contained in equations (6) and (7) agrees with the number found in equations (3a)-(3d).

6 Show that

$$\frac{\partial \mathcal{F}^{\mu\nu}}{\partial x^\lambda} + \frac{\partial \mathcal{F}^{\nu\lambda}}{\partial x^\mu} + \frac{\partial \mathcal{F}^{\lambda\mu}}{\partial x^\nu} = 0$$

(equation (7), p. 234) implies that the magnetic field has zero divergence.

7 Write down the fields of an electromagnetic plane wave propagating in the z direction, choosing some polarization. Do not assume a sinusoidal wave. Show that this is a solution of

$$\frac{\partial \mathcal{F}^{\mu\nu}}{\partial x^\nu} = 0$$

(equation (6), p. 234, in a vacuum).

Photo Credits

15 *Atomic clock on plane*: Copyright 1971, Associated press, used under U.S. fair use exception to copyright law. **18** *Ring laser gyroscope*: Wikimedia commons user Nockson, CC-BY-SA licensed. **21** *Machine gunner's body*: Redrawn from a public-domain photo by Cpl. Sheila Brooks. **21** *Machine gunner's head*: Redrawn from a sketch by Wenceslas Hollar, 17th century. **21** *Minkowski*: From a 1909 book, public domain. **29** *Muon storage ring at CERN*: Copyright 1974 by CERN; used here under the U.S. fair use doctrine. **30** *Joan of Arc holding banner*: Ingres, 1854. **30** *Joan of Arc interrogated*: Delaroche, 1856. **35** *Surveyors*: Don Swanson, public domain; Owais Khursheed, CC-BY-SA; Paul Peterson, public domain; British Royal Navy, public domain. **35** *Pigeon*: Wikimedia commons user nitramtrebla, CC-BY-SA. **35** *Mushroom cloud*: USGS, public domain. **84** *Photo of PET scanner*: Wikipedia user Hg6996, public domain. **84** *Ring of detectors in PET scanner*: Wikipedia user Damato, public domain. **84** *PET body scan*: Jens Langner, public domain. **85** *Oscilloscope trace*: From Bertozzi, 1964; used here under the U.S. fair use doctrine. **95** *Gamma-Ray burst*: NASA/Swift/Mary Pat Hrybyk-Keith and John Jones. **92** *Gamma-ray spectrum*: Redrawn from a public-domain image by Kieran Maher and Dirk Hunniger. **120** *Eotvos*: Unknown source. Since Eötvös died in 1919, the painting itself would be public domain if done from life. Under U.S. law, this makes photographic reproductions of the painting public domain. **121** *Artificial horizon*: NASA, public domain. **122** *Pound and Rebka photo*: Harvard University. I presume this photo to be in the public domain, since it is unlikely to have had its copyright renewed. **136** *Surfer*: Redrawn from a photo by Jon Sullivan, CC0 license. **147** *Lambert projection*: Eric Gaba, CC-BY-SA. **154** *Levi-Civita*: Believed to be public domain. Source: <http://www-history.mcs.st-and.ac.uk/PictDisplay/Levi-Civita.html>.

Index

- aberration, 134
- abstract index notation, 139
 - equivalent to birdtracks, 139
- acceleration
 - proper, 59, 61
 - vector, 61
- affine parameter, 152
- angular momentum, 165
 - conservation of, 193
- Bell's spaceship paradox, 71, 206
- Bell, John, 71
- birdtracks, 126
- birdtracks notation, 126
 - equivalent to abstract index notation, 139
- black hole, 107
- Bohr model, 172
- boost
 - defined, 22
- Born rigidity, 203
- Bridgman, P.W., 26
- Brown-Bethe scenario, 107
- Cauchy surface, 158
- Cauchy-Schwarz inequalities, 36
- causality, 15, 44
- caustic, 204
- center of mass frame, 89
- Chandrasekhar limit, 105
- Christoffel symbol, 195
- clock-comparison experiments, 125
- collision
 - invariants, 89
- congruence, 202
- coordinates
 - Minkowski, 20
- correspondence principle
 - defined, 15
 - for time dilation, 15
- covariant derivative, 151, 193
 - in relativity, 193
- covector, 127
- curl, 232
- current vector, 175
- Cvitanović, Predrag, 126
- de Sitter, Willem, 172
- degree of freedom, 62
- derivative
 - covariant, 193
 - in relativity, 193
- diffeomorphism, 127
- divergence, 178, 232
- Doppler shift, 55, 133
- duality, 126, 127, 130
- dust, 180
- Eötvös experiments, 120
- Einstein summation convention, 139
- Einstein synchronization, 19
- electron capture, 106
- energy
 - equivalent to mass, 82
- event, 12
- event horizon, 62
 - black hole, 107
- expansion scalar, 203
- fine structure constant, 20, 166
- force, 100
 - four-vector, 100
 - three-vector, 100
- four-vector, 60
- gamma factor
 - as an inner product, 64
 - defined, 25
- gamma ray
 - pair production, 93
- garage paradox, 33
- Gauss's theorem, 188
- geodesic, 195, 197
 - differential equation for, 198
- geodesic equation, 199
- Goudsmit, 172
- group velocity, 137
- Herglotz-Noether theorem, 205
- horizon, 62

- hyperbolic motion, 61, 76
- inner product, 24
- interval
 - defined, 24
- invariant
 - compared to scalar, 127, 128
 - defined, 22
- Ives-Stilwell experiments, 56
- Lambert cylindrical projection, 147
- Levi-Civita tensor, 154, 224
- Levi-Civita, Tullio, 182
- Lewis-Tolman paradox, 67
- light cone, 17
- Lorentz invariance, 47
- Lorentz invariant, *see* invariant
- Lorentz scalar, *see* scalar
- Lorentz transformation, 30
- lowering an index, 141
- mass
 - defined, 88
 - equivalent to energy, 82
- metric, 24
- Minkowski coordinates, 20
- Minkowski, Hermann, 20
- natural units, 24
- neutron star, 105, 106
- normalization, 63
- operationalism, 26
- orientation, 153
- pair production, 93
- paradox
 - Bell's spaceship, 71, 206
- parallel transport, 45
- Penrose
 - graphical notation for tensors, 126
- phase velocity, 137
- polarization
 - of light, 94
- positron, 84
- projection operator, 66
- proper acceleration, 61
- proper time, 23
- pulsar, 106
- raising an index, 141
- rank of a tensor, 147
- rapidity, 59
- rigidity, 203
- Rindler coordinates, 144
- scalar, 127, 128
 - compared to Lorentz invariant, 127
- signature, 24
- spaceship paradox, 71, 206
- stress-energy tensor, 179
 - interpretation of, 183
- surfaces
 - Cauchy, 158
 - classification of, 158
- tensor
 - Penrose graphical notation, 126
 - rank, 147, 180, 208
- thickening of a curve, 70
- Thomas precession, 172
- Thomas, Llewellyn, 172
- three-vector, 67
- Tolman-Oppenheimer-Volkoff limit, 107
- topology, 47, 154
- torsion, 196
- transverse polarization
 - of light, 94
- triangle inequalities, 36
- tubular neighborhood, 73
- twin paradox, 25
 - Galilean, 14
 - signals exchanged, 55
- Uhlenbeck, 172
- units
 - for tensors, 207
- vector
 - distinguished from covector, 127
 - Penrose graphical notation, 126
- velocity
 - addition, 57
 - vector, 60
 - wave
 - group, 137
 - phase, 136
- volume
 - 3-volume covector, 157

affine, 151
Voyager space probe, 38

Waage, Harold, 117
wavenumber, 129
Wheeler, John, 118
white dwarf, 105
work, 102
world-line, 12

Yukawa potential, 97